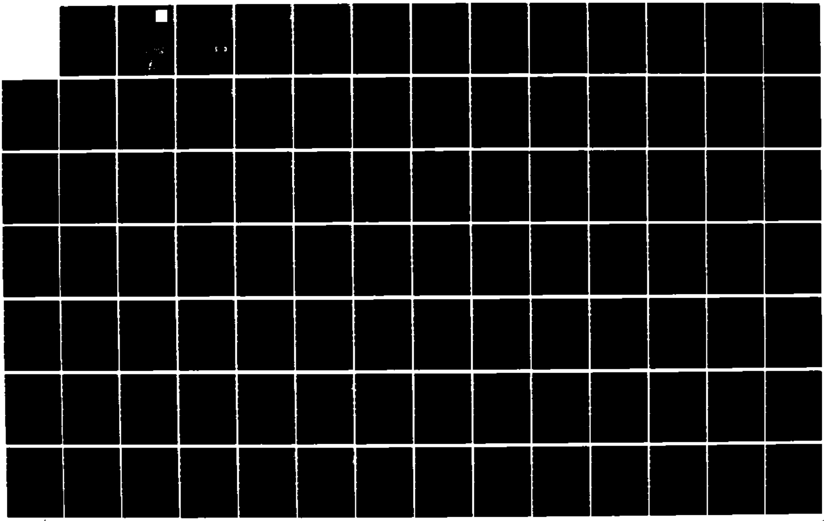
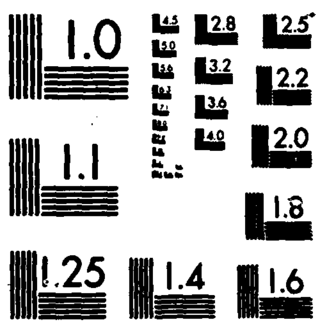
AD-A163 642 THE CALCULUS OF UNCERTAINTY IN ARTIFICIAL INTELLIGENCE 1/3
AND EXPERT SYSTEMS..(U) GEORGE WASHINGTON UNIV
WASHINGTON DC INST FOR RELIABILITY AND..
UNCLASSIFIED N D SINGPURWALLA ET AL. 15 JAN 86 F/G 9/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

THE CALCULUS OF UNCERTAINTY IN
ARTIFICIAL INTELLIGENCE AND
EXPERT SYSTEMS*

Proceedings of a Conference
held on December 28-29, 1984

THE
GEORGE
WASHINGTON
UNIVERSITY

AD-A163 642

STUDENTS FACULTY STUDY R
ESEARCH DEVELOPMENT FUT
URE CAREER CREATIVITY CC
MMUNITY LEADERSHIP TECH
NOLOGY FRONTIE SIGN
ENGINEERING APP ENC
GEORGE WASHIN NIV

SCHOOL OF ENGINEERING
AND APPLIED SCIENCE

THE CALCULUS OF UNCERTAINTY IN ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEMS*

Proceedings of a Conference
held on December 28-29, 1984

GWU/IRRA/Serial TR-86/2

The George Washington University
School of Engineering and Applied Science
Institute for Reliability and Risk Analysis

*This work was supported by the National Security Agency under Grant N00014-85-G-0162 issued by the Office of Naval Research, and by Contract N00014-85-K-0202 with The George Washington University.

CONTENTS

Foreword

Despite the remarkable progress in the use and application of artificial intelligence and expert systems techniques in the past ten years, several fundamental issues remain unresolved.

One of these is how best to deal with uncertainty in the conditions of interest involving the use of expert systems. Even with the increased pace of discovery and innovation in the mathematical and information sciences, there still remain to be resolved issues pertaining to methods adequate for the treatment of uncertainty which are acceptable to all practitioners. Obviously many philosophical and methodological questions need to be addressed.

It was clearly recognized by researchers in government and universities that a conference to address these issues and to at least focus some of the thoughts of scientists, develop awareness and concern, and share experiences would be a worthwhile happening. In particular, the scientific and technical interests of the Office of Naval Research and the National Security Agency were important factors in motivating the organization of such a meeting. Edward J. Wegman of ONR and C. Terrance Ireland and James Maar of NSA were responsible for stimulating the concept and fostering its realization. Professor Nozer D. Singpurwalla, Professor of Operations Research and of Statistics at the George Washington University, was the key to accomplishing the transformation from idea to reality, and was the organizer and driving force to implement this common aspiration.

Accordingly, plans were made to convene a conference, the first of its kind, at the George Washington University. The GWU Institute for Reliability and Risk Analysis would host the event and also would be able to provide several key participants who are experts in the subject areas. Good fortune had arranged for Professor Dennis V. Lindley and Stephen R. Watson of Cambridge University to be visiting at the Institute, and the Department of Operations Research, respectively, at the appropriate time. The agencies provided sponsorship, and plans moved forward. The conference would be limited to a small number of researchers and interested practitioners.

The meeting was deliberately restrained to be a low key event and took place on December 28 and 29, 1984, at a time when the University was closed for the winter holiday period. Nevertheless, the University interest was not to be diminished. President Lloyd H. Elliott played a personal role in getting the conference off to a good start by providing a thoughtful and witty talk which not only launched the meeting in an intellectually stimulating manner but also clearly demonstrated the University's support for the conference and its subject.

This report is an attempt to document the proceedings of the conference in a manner that will provide a record of what transpired for the sponsors and participants, and also provide resource material for those interested in the topic from any of several perspectives.

It is hoped and expected that the original material prepared for the conference will eventually be published in the open scientific literature. Now, however, this is a record of the actual presentations and discussions (or as close as we could get to it) by the participants as well as the original technical papers. The discussions were edited and smoothed only very lightly, as will be apparent from a cursory reading. As was to be expected, not all of the participants provided uniform inputs to facilitate the documentation task.

A genuine shortcoming of these proceedings is the inability to completely capture and reproduce the effectiveness of the speaker's use of the overhead projector. We have included some of those slides as selected by the speakers; however much of the communication was due to the vigor of the presentations and audience interactions, frequently making use of spontaneous but valuable on the spot transparencies, some of which unfortunately cannot be reproduced here.

A most important factor in ensuring the success of the meeting was that of the role played by the Moderator, Professor Morris DeGroot of the Department of Statistics, Carnegie-Mellon University. His conduct of the meeting during its two days, and performance as a catalytic agent, interpreter, and clairvoyant was most important to the outcome. It is truly felt that the transcript alone cannot adequately reflect his essential contributions in this regard. To those who were present, it was easily recognized as the privilege of witnessing a most enjoyable tour de force by a scientist who is uniquely talented and expertly knowledgeable, and generous with his ready wit and humor.

Grateful acknowledgment is made of the assistance of Professors Donald Gross and Graham W. McIntyre of the GWU School of Engineering and Applied Science in solving critical administrative problems incident to the meeting, and to Mrs. Teresita R. Abacan in typing this report.

Seymour M. Selig
Coordinating Editor

KEY PARTICIPANTS

Invited Speakers (in order of presentation):

Glenn Shafer, School of Business
University of Kansas
Laurence, KS

Lotfi A. Zadeh, Computer Science Division
University of California, Berkeley
Berkeley, CA

Dennis V. Lindley, The George Washington University
(Visiting), Washington, DC

David Speigelhalter, Medical Research Council Centre
Cambridge, England

Invited Discussants:

Arthur P. Dempster, Department of Statistics
Harvard University
Cambridge, MA

Stephen R. Watson, The George Washington University
(Visiting), Washington, DC

Moderator: Morris H. DeGroot, Department of Statistics
Carnegie-Mellon University, Pittsburgh, PA

ATTENDANCE LIST

CONFERENCE ON THE CALCULUS OF UNCERTAINTY IN ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEMS

December 27-28, 1984

PATRICK BAILEY
Naval Oceans System Center
271 Catalina Blvd.
San Diego, CA 92152

LEE S. BROWNSTON
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213

MARVIN S. COHEN
Decision Science Consortium, Inc.
7700 Leesburg Pike, Suite 421
Falls Church, VA 22043

S. G. CORTELYOU
SEAS, CEEP
George Washington University
Washington, DC 20052

MORRIS H. DEGROOT
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213

ARTHUR P. DEMPSTER
Department of Statistics
Harvard University
Cambridge, MA 02138

LLOYD H. ELLIOTT
President
George Washington University
Washington, DC 20052

PAUL S. FISCHBECK
Operations Research Department
Naval Postgraduate School
Monterey, CA 93943

DONALD GROSS
Department of Operations Research
George Washington University
Washington, DC 20052

HENRY HAMBURGER
Naval Research Laboratory
Washington, DC 20375

HERBERT H. HOLMAN
Department of Defense
Ft. George G. Meade MD 20755

C. TERRANCE IRELAND
National Security Agency
Ft. George G. Meade MD 20755

JOHN KAY
Department of Engineering
Administration
George Washington University

GABRIEL PEI
IBM Federal Systems Division
9500 Godwin Drive
Manassas, VA 22110

ARTHUR D. KIRSCH
Dept. of Statistics/
Computers Info. Systems
George Washington University
Washington, DC 20052

RICHARD Y. PEI
The Rand Corporation
2100 M St., N.W.
Washington, DC 20037

AUGUSTINE KONG
Harvard University
Statistics Department
One Oxford Street, 6th Floor
Cambridge, MA 02138

JOHN D. PRANGE
Department of Defense
Ft. George G. Meade MD 20755

JAY LIEBOWITZ
Department of Management
Science
George Washington University
Washington, DC 20052

PHILIP N. REEVES
Department of Health
Services Adm.
George Washington University
Washington, DC 20052

DENNIS V. LINDLEY
Department of Operations
Research
George Washington University
Washington, DC 20052

SEYMOUR M. SELIG
Institute for Reliability
and Risk Analysis
George Washington University
Washington, DC 20052

JAMES RICHARD MAAR
National Security Agency
Ft. George G. Meade MD 20755

GLENN SHAFER
School of Business
University of Kansas
Lawrence, Kansas 66045

ALAN L. MEYROWITZ
Office of Naval Research
Information Sciencies Division
800 N. Quincy Street
Arlington, VA 22217

J. RANDOLPH SIMPSON
Office of Naval Research
Arlington, VA 22217

NOZER D. SINGPURWALLA
Department of Operations
Research
George Washington University
Washington, DC 20052

STEPHEN R. WATSON
Cambridge University
Engineering Dept.
Control and Management
Systems Div.
Mill Lane
Cambridge, CB2 IRX England

ROBERT SMYTHE
Department of Statistics
George Washington University
Washington, DC 20052

EDWARD J. WEGMAN
Office of Naval Research
800 N Quincy Street
Arlington, VA 22217

HENRY SOLOMON
Graduate School of Arts &
Sciences
George Washington University
Washington, DC 20052

BEN P. WISE
Carnegie-Mellon University
Robotics Institute and Dept.
of Engineering and Public
Policy
Schenley Park
Pittsburgh, PA 15213

RICHARD M. SOLAND
Department of Operations
Research
George Washington University
Washington, DC 20052

RONALD R. YAGER
Machine Intelligence
Institute
Iona College
New Rochelle, NY 10801

REFIK SOYER
Department of Operations
Research
George Washington University
Washington, DC 20052

LOTFI A. ZADEH
Computer Science Division
University of California
Berkeley, CA 94720

DAVID SPEIGELHALTER
MRC Biostatistics Unit
Medical Research Council Centre
Hills Road
Cambridge, CB2 2QH England

PROBABILITY JUDGMENT IN ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEMS

Glenn Shafer
School of Business
University of Kansas

CONTENTS

This paper was prepared for the Conference on the Calculus of Uncertainty in Artificial Intelligence and Expert Systems held at George Washington University, December 27 and 28, 1984. Research for the paper was partially supported by NSF grant IST-8405210.

I have been asked to speak on the use of belief functions in artificial intelligence and expert systems. For the sake of perspective, I propose to address the broader topic indicated by my title. The theory of belief functions is part of the theory of probability judgment, and a general understanding of the role of probability judgment in artificial intelligence can help us understand the particular role of belief functions.

I will not attempt to evaluate all the ways in which probability has been used in artificial intelligence, nor even all the ways in which belief functions have been used. Instead, I will aim for some general insights into the interaction between probability ideas and artificial intelligence ideas. Many of my comments will be historical. I hope readers will forgive me for those cases where I belabor the obvious or repeat the well-known; my excuse is that I hope to reach a dual audience--students of probability who may not know very much about artificial intelligence, and students of artificial intelligence who may not know very much about probability.

The first two sections of the paper are introductory in nature. Section 1 considers the reasons for the artificial intelligence community's initial disinterest in probability and its recent change of heart and outlines the paper's conclusions about the how current expert systems fall short of putting probability judgment into artificial intelligence. Section 2 deals with probability judgment without reference to artificial intelligence; here I discuss the split between Bayesian and non-

Bayesian methods and place the theory of belief functions in this historical context.

Sections 3 studies some strands of the development within artificial intelligence of ideas about using probability judgment in expert systems. Here we see how the general issues that separate the Bayesian and belief-function theories appear in the context of expert systems, and we gain some insight into why flexibility is harder to achieve with probability judgment than with other kinds of reasoning. Section 4 discusses the problem of giving an artificial intelligence a genuine capacity for probability judgment.

1. The Emergence of Probability in Artificial Intelligence

Until recently, the artificial intelligence community showed relativly little interest in probability. There is little probabilility, for example, in the three volume Handbook of Artificial Intelligence, published in 1981 and 1982. During the past two or three years, however, probability and the management of uncertainty in intelligent systems has become a widely discussed topic. Why the initial disinterest, and why the change?

The reasons for the initial disinterest are clear. Probabilities are numbers, and number crunching is just what artificial intelligence was supposed not to be. When the artificial intelligence community was founded, computers were used mainly for number crunching. They were impressively good at this, but they were not intelligent. Intelligence seemed to require more general kinds of symbol manipulation.

Moreover, when we begin to think about computer programs

that will match the achievements of human intelligence, we find that we are thinking about programs with non-numerical inputs and outputs. What place is there for talk about numbers in the case of these programs? They are merely sets of rules for going from the inputs to the outputs, and while it might be possible to identify some intermediate steps that are analogous to operations on numerical probabilities, it seems pointless to do so. It seems better to tell what is really going on.

The prejudice against numbers in general and probabilities in particular has not entirely disappeared from artificial intelligence, and the argument sketched in the preceding paragraph is still made. Paul Cohen and Jon Doyle made it during the panel discussion on uncertainty at the meeting of the American Association for Artificial Intelligence in Austin last summer. Cohen went on to argue that probability talk should be replaced by talk about reasons and endorsements--we should spell out what endorsements a program requires before it will take a given action or draw a given conclusion (Cohen, 1983). Doyle argued that the problem of combining uncertain evidence should be solved not by numerical calculations but by the techniques of non-monotonic logic (Doyle, 1979).

But the factors that caused this prejudice have substantially changed. The vague idea that artificial intelligence can be defined largely through the contrast with number crunching has been replaced by the equally vague but equally powerful idea that intelligence is produced by complexity and by access to large amounts of knowledge. And two specific openings have appeared

for probability:

(1) The absolute ban on non-numerical inputs has been dropped. In addition to programs that try to match aspects of human intelligence, artificial intelligence is now also concerned with expert systems and other intelligent systems that interact with human users and can use numerical inputs supplied by these users.

(2) The artificial intelligence community has absorbed David Marr's views on levels of explanation. In his work on vision, Marr convincingly made the point that full understanding of an intelligent system involves explanation at various levels. In addition to explanation at the level of implementation (what is really going on) we also need explanation at more abstract levels. "It's no use, for example, trying to understand the fast Fourier transform in terms of resistors as it runs on an IBM 370." (Marr, 1982, p. 337) Understanding of this point takes the rhetorical force out of the argument that there is no place for probability ideas when inputs and outputs are non-numerical.

Most of the current interest in probability in artificial intelligence is the result of opening (1). In many areas it impossible to build expert systems without the use of probability. But I will argue in this paper that opening (2) is a more genuine opening for probability in artificial intelligence. Because of (2), we can now recognize the value to an artificial intelligence of an ability to design probability arguments and generate the numerical judgments they require.

The ban on numerical inputs in artificial intelligence was

dropped because the artificial intelligence community became interested in expert systems. Why did this happen? The answer is that the community discovered ways of building expert systems that incorporated ideas that seemed to reflect important aspects of human intelligence. As I explain in section 3 below, most of the expert systems developed within artificial intelligence have been production systems, and production systems seem to have the flexibility in acquiring and using knowledge that is characteristic of intelligence.

I argue in this paper that the expert systems we can now build to use probability judgments do not have this kind of flexibility and hence should not be classed under the heading of artificial intelligence. The problem seems to be that probability judgment requires an overall design and hence cannot be achieved by relatively unstructured methods of programming applied to unstructured probability judgments.

As a result of the explosion of interest in expert systems, the field of artificial intelligence is now struggling to maintain its sense of identity. The idea of an expert system began in artificial intelligence, but any system with expert capabilities can justifiably claim the name, whether it is written in LISP or FORTRAN, and many systems developed outside of artificial intelligence have more impressive expert capabilities than those developed inside it. It is clear, therefore, that artificial intelligence must withdraw from its embrace of the whole field of expert systems in order to maintain intellectual coherence. But it is unclear just what parts of the field of expert systems will

remain in the embrace. My suggestion here is that artificial intelligence will retain its newfound interest in probability but will look beyond the current expert systems to deeper uses of probability ideas.

2. Bayesian and Belief-Function Arguments

In this section I review some general ideas about probability judgment, without reference to the particular problems of artificial intelligence. I begin by sketching a way of looking at the frequentist vs. Bayesian controversy, a controversy that has dominated discussions of probability judgment for more than a century. After developing a constructive understanding of the Bayesian theory, I introduce another constructive theory, the theory of belief functions. I argue that both theories should be thought of as languages for expressing probability judgments and constructing probability arguments.

2.1. Two Strategies for Probability Judgment. What we now call the mathematical theory of probability was originally called the theory of games of chance. Probability was an entirely different topic; something was probable when there was a good argument or good authority for it. When James Bernoulli and others began to use the word probability in connection with the theory of games of chance, they were expressing the ambition that this theory might provide a general framework for evaluating evidence and weighing arguments. But just how might this work? How can the theory of games of chance help us evaluate evidence?

In the nineteenth century, it became clear that there are

two distinct strategies for relating evidence to the picture of chance. Today, these two strategies might be called the frequentist and Bayesian strategies, but in order to avoid some of the connotations of these names, let me call them, for the moment, the direct probability and conditional probability strategies.

The direct probability strategy relies on direct application of the idea that in life, as in games of chance, what happens most often is most likely to happen in a particular case under consideration. The ideal kind of evidence for this strategy is knowledge of the frequency of outcomes in similar cases. I assign a 98% probability to the prediction that a student who first appears three weeks after the beginning of my elementary statistics course will not be able to pass the course, because it has almost always turned out that way in the past.

The conditional probability strategy uses the picture of chance in a deeper way. It observes that games of chance unfold step-by-step, with the probabilities for different possible final outcomes changing at each step, and it suggests that the accumulation of evidence should change probabilities in a similar step-by-step way. Thus my probability for whether the late-appearing student will pass my course should change when I learn more about his history and circumstances, just as my probability for whether two successive rolls of a die will add to nine will change when I learn the result of the first roll. The conditional probability strategy usually leads to a more complicated argument than the direct probability strategy, since it involves construction of a probability measure over a more complicated frame and then the reduction of this measure and frame by conditioning.

In general, there is not, I believe, any _a priori_ reason to prefer one of these two stategies to the other. We cannot say that it is normative to use one and irrational to use the other. They are both stategies for producing arguments, and it is the arguments that must be evaluated as convincing or unconvincing. It may be most convincing to lump this late-appearing student with all my past late-appearing students, with the general excuse that particulars have not made much difference in the past. Or I may have had enough experience with late-appearing students like this one on some particulars that I can make a more convincing direct probability argument by looking at the past frequency of success just for these late-appearing students. Or, on the other hand, I may have the experience and insight needed to convincingly make probability judgments from which I can construct a probability measure that I can condition on the particulars. The issue cannot be settled in the abstract, without reference to the experience I bring to bear on the problem.

I also believe that neither of the two strategies is inherently more objective or subjective than the other. It is true that the direct probability strategy, since it tends to consider broader classes, is more likely to result in probability judgments based on actual frequency counts. But the objectivity of these frequencies must always be coupled with a subjective judgment of their relevance. And even with broad classes we most often have hunches and impressions rather than actual counts.

Historically, however, the direct probability strategy has come to be associated with claims to objectivity, while the

conditional probability approach has come to the associated with claims to rationality. This fact seems to be a result of efforts to square the interpretation of probability with the empiricist and positivist philosophical trends of the late nineteenth and early twentieth centuries.

2.2. The Frequentist vs. Bayesian Deadlock. Laplace, writing at the beginning of the nineteenth century, was able to define numerical probability as the measure of the "reason we have to believe." But by the middle of the nineteenth century, many students of probability were looking for a more empirical definition. They found this definition in the idea of frequency, and they proceeded to reject those applications of probability theory that could not be based on observed frequencies. In particular, they rejected Laplace's method of calculating the probability of causes, which is a special case of the conditional probability strategy.

The frequentist philosophy severely restricted the domain of application of numerical probability, and those who wanted to use numerical probability more generally were forced to search for a philosophical foundation for the conditional probability strategy that would fit the positivist mind-set. Such a philosophical foundation was finally established in the twentieth century by Ramsey, de Finetti, and especially Savage. These authors conceived the idea that sujective probability should be given a behavioral and hence positivist interpretation--a person's probabilities should be derivable from his choices. They formulated postulates for what they called rational behavior, postulates

which assure that a person's choices do determine numerical probabilities. And they argued that it is normative to follow these postulates and hence normative to have subjective probabilities.

During the past two decades, the philosophical foundation provided by Savage's postulates has led to a remarkable resurgence, both mathematical and practical, of the conditional probability strategy. The resulting body of theory has been called "Bayesian," because the conditional probability strategy often uses Bayes's theorem.

Though the new Bayesian philosophy has played a historically valuable role in rescuing the conditional probability strategy from its frequentist opponents, it has its own obvious shortcomings. Most important, perhaps, is its inability to explain how the quality of a probability analysis depends on the availability and quality of relevant evidence. Whereas the frequentist philosophy tries to limit applications of probability to models for which we have clearly relevant and objective frequency counts, there is nothing in the Bayesian philosophy to make our choice of a model depend in any way on the availability of relevant evidence. The postulates apply equally to any model.

We have, then, a deadlock between two inadequate philosophies of probability. On the one side, the frequentist philosophy, which recognizes the relevance of evidence but tries to justify claims to objectivity by limiting numerical probability judgment to cases where the evidence is of an ideal form; on the other side, the Bayesian philosophy, which recognizes the subjectivity of all probability judgment but ignores the quality of

evidence and claims it is normative to force all probability judgment into one particular mold.

We have been caught in this deadlock for three decades. We have tired of it, and we are inclined to ask the two sides to compromise (see, e.g., Box, 1980). But we have not been able to find a philosophical foundation for probability judgment that can resolve the deadlock.

I believe that the way out of the deadlock is to back up and recognize that a positivist philosophical account of probability is no longer needed. Our intellectual culture has moved away from positivism and towards various sorts of pragmatism, and once we recognize this we will be free to discard both the frequentists' claims to objectivity and the Bayesians' claims to normativeness.

2.3. Constructive Probability. In several recent papers (especially Shafer, 1981, and Shafer and Tversky, 1985) I have proposed the name "constructive probability" for the pragmatic, post-positivist foundation that I think we need for probability judgment. The idea is that numerical probability judgment involves fitting an actual problem to a scale of canonical examples. The canonical examples usually involve the picture of chance in some way, but different choices of canonical examples are possible, and these different choices provide different theories of subjective probability, or, if you will, different languages in which to express probability judgments. No matter what language is used, the judgments expressed are subjective; the subjectivity enters when we judge that the evidence in our actual

problem matches in strength and significance the evidence in the canonical example.

Within a given language of probability judgment, there can be different strategies for fitting the actual problem to the scale of canonical examples. The direct and conditional probability strategies described above live, I think, in the same probability language, the language in which evidence about actual questions is fit to canonical examples where answers are determined by known chances. We may call this language the Bayesian language. (For a more detailed account of different strategies that are available within the Bayesian language, see Shafer and Tversky, 1985. The distinction between the direct and conditional probability strategies corresponds to the distinction that is made there between total-evidence and conditioning designs.)

The constructive viewpoint tells us that when we work within the Bayesian language we must make a judgment about how far to take the conditional probability strategy in each particular problem. And we make this judgment on the basis of the availability of evidence to support the conditional and unconditional probability judgments that are required.

It may be useful to elaborate this point. Suppose we want to make probability judgments about a frame of discernment S. (A <u>frame of discernment</u> is a list of possible answers to a question; so this means we want to make probability judgments about which answer is correct.) We reflect on what relevant evidence we have, and produce a list $E_1,\ldots,E_n$ of facts that seem to summarize this evidence adequately. The conditional probability

strategy amounts to standing back from our knowledge of these n facts, pretending that we did not yet know them, and constructing a probability measure over a frame that considers not only the question considered by S but also the question whether $E_1,\ldots,E_n$ are or are not true; typically we construct this measure by making probability judgments P(s) and $P(E_1\&\ldots\&E_n|s)$ for each s in S. The problem with this strategy is that we now need to look for evidence on which to base these probability judgments. We have used our best evidence up, as it were, but now we have an even larger judgmental task than before. According to the behaviorist Bayesian theory, there is no problem--it is normative to have the requisite probabilities, whether we can identify relevant evidence or not. But according to the constructive viewpoint, there is a problem, a problem which limits how far we want to go. We may want to apply the conditional probability strategy to some of the E_i, but we may want to reserve the others to help us make the probability judgments (see Shafer and Tversky, 1985).

2.4. The Language of Belief Functions. Whereas the Bayesian probability language uses canonical examples where known chances are attached directly the possible answers to the question asked, the language of belief functions uses canonical examples where known chances may be attached only to the possible answers to a related question.

Suppose, indeed, that S and T denote, respectively, the possible answers to two distinct but related questions. When we say that these questions are related, we mean that a given answer to one of the questions may not be compatible with all the possi-

ble answers to the other. Let us write "sCt" when s is an element of S, t is an element of T, and s and t are compatible. Given a probability measure P over S (I assume for simplicity that P is defined for all subsets of S), we may define a function Bel on subsets of T by setting

$$\mathrm{Bel}(B) = P\{s \mid \text{if } sCt, \text{ then } t \text{ is in } B\}. \qquad (1)$$

for each subset B of T. The right-hand side of (1) is the probability that P gives to those answers to the question considered by S that require the answer to the question considered by T to be in B; the idea behind (1) is that this probability should be counted as reason to believe that the latter answer is in B. We might, of course, have more direct evidence about the question considered by T, but if we do not, or if we want to leave other evidence aside for the moment, then we may call Bel(B) a measure of the reason we have to believe B based just on P.

I call the function Bel given by (1) the belief function obtained by extending P from S to T. A probability measure P is a special kind of belief function; this is just the case where (i) S=T and (ii) sCt if and only if s=t.

All the usual devices of probability are available to the language of belief functions, but in general they are applied in the background, at the level of S, before extending to degrees of belief on T, the frame of interest. Thus the language of belief functions is a generalization of the Bayesian language. I have studied the language of belief functions in detail in earlier work--see especially Shafer (1976,1985). Here I will use some examples of (1) to illustrate the language and to contrast it with the Bayesian language.

Example 1. Is Fred, who is about to speak to me, going to speak truthfully, or is he, as he sometimes does, going to speak carelessly, saying something that comes into his mind, but the truth of which he does not know? Let S denote the possible answers to this question; S={truthful,careless}. Suppose I know from experience that Fred's announcements are truthful reports on what he knows about 80% of the time and are careless statements the other 20% of the time. Then I have a probability measure P over S: P{truthful}=.8, P{careless}=.2.

Are the streets outside slippery? Let T denote the possible answers to this question; T={yes,no}. And suppose Fred's announcement turns out to be, "The streets outside are slippery." Taking account of this, I have a compatibility relation between S and T; "truthful" is compatible with "yes" but not with "no," while "careless" is compatible with both "yes" and "no." Applying (1), I find

$$\mathrm{Bel}(\{\text{yes}\})=.8 \quad \text{and} \quad \mathrm{Bel}(\{\text{no}\})=0; \tag{2}$$

Fred's announcement gives me an 80% reason to believe that the streets are slippery outside, but no reason to believe that they are not.

How might a Bayesian argument using this evidence go? The direct probability strategy would use all my evidence, Fred's announcement included, to make a direct probability judgment about whether the streets are slippery. But if I want an argument that uses the judgment that Fred is 80% reliable as one ingredient, then I will use a conditional probability strategy. This strategy requires two further probability judgments: (1) A

prior probability, say p, for the proposition that the streets are slippery; this will be a judgment based on evidence other than Fred's announcement. (2) A conditional probability, say q, that Fred's announcement will be accurate even though it is careless. Given these ingredients, I can calculate a Bayesian probability that the streets are slippery given Fred's announcement and my other evidence:

$$P(\text{slippery}|\text{announcement}) = \frac{.8p + .2pq}{.8p + .2pq + .2\,(1-p)(1-q)}. \qquad (3)$$

Is the Bayesian argument (3) better than the belief-function argument (2)? This depends on whether I have the evidence required. If I do have evidence to support the judgments p and q--if, that is to say, my situation really is quite like a situation where the streets and Fred are governed by known chances, then (3) is a good argument, clearly more convincing than (2) because it takes more evidence into account. But if the evidence on which I base p and q is of much lower quality than the evidence on which I base the number 80%, then (2) will be more convincing.

The traditional debate between the frequentist and Bayesian views has centered on the quality of evidence for prior probabilities. It is worth remarking, therefore, that we might well feel that q, rather than p, is the weak point in the argument (3). I probably will have some other evidence about whether it is slippery outside, but I may not have any idea about how likely it is that Fred's careless remarks will accidentally be true.

A critic of the belief-function argument (2) might be tempted to claim that the Bayesian argument (3) shows (2) to be wrong even if I do lack the evidence needed to supply p and q.

Formula (3) gives the correct probability for whether the street is slippery , the critic might contend, even if I cannot say what this probability is, and it is almost certain to differ from (2). This criticism is fundamentally misguided. In order to say that (3) gives the "correct" probability, I must be able to convincingly compare my situation to the picture of chance. And my inability to model Fred when he is being careless is not just a matter of not knowing the chances--it is a matter of not being able to fit him into a chance picture at all.

Example 2. Suppose I do have some other evidence about whether the streets are slippery: my trusty indoor-outdoor thermometer says that the temperature is 31^0 Fahrenheit, and I know that because of the traffic ice could not form on the streets at this temperature.

My thermometer could be wrong. It has been very accurate in the past, but such devices do not last forever. Suppose I judge that there is a 99% chance that the thermometer is working properly, and I also judge that Fred's behavior is independent of whether it is working properly or not. (For one thing, he has not been close enough to my desk this morning to see it.) Then I have determined probabilities for the four possible answers to the question, "Is Fred being truthful or careless, and is the thermometer working properly or not?" For example, I have determined the probability .8x.99=.792 for the answer "Fred is being truthful, and the thermometer is working properly." All four possible answers, together with their probabilities, are shown in the first two columns of Table 1. We may call the set of these

four answers our new frame S.

Taking into account what Fred and the thermometer have said, I have the compatibility relation between S and T given in the last column of the table. (Recall that T considers whether the streets are slippery; T={yes,no}.) The element (truthful,working) of S is ruled out by this compatibility relation (since Fred and the thermometer are contradicting each other, they cannot both be on the level); hence I condition the initial probabilities by eliminating the probability for (truthful,working) and renormalizing the three others. The resulting posterior probabilities on S are given in the third column of the table.

Finally, applying (1) with these posterior probabilities on S, I obtain the degrees of belief

$$\mathrm{Bel}(\{\text{yes}\})=.04 \quad \text{and} \quad \mathrm{Bel}(\{\text{no}\})=.95. \tag{4}$$

This result reflects that fact that I put much more trust in the thermometer than in Fred.

The preceding calculation is an example of <u>Dempster's rule of combination</u> for belief functions. Dempster's rule combines two or more belief functions defined on the same frame but based

s	Probability of s Initial	Probability of s Posterior	Elements of T compatible with s
(truthful,working)	.792	0	--
(truthful,not)	.008	.04	yes
(careless,working)	.198	.95	no
(careless,not)	.002	.01	yes,no

Table 1.

on independent arguments or items of evidence; the result is a belief function based on the pooled evidence. In this case the belief function given by (2), which is based on Fred's testimony alone, is being combined with the belief function given by

$$\text{Bel}(\{\text{yes}\})=0 \quad \text{and} \quad \text{Bel}(\{\text{no}\})=.99, \tag{5}$$

which is based on the evidence of the thermometer alone. In general, as in this example, Dempster's rule corresponds to the formation and subsequent conditioning of a product measure in the background. See Shafer (1985) for a precise account of the independence conditions needed for Dempster's rule.

Example 3. Dempster's rule applies only when two items of evidence are independent, but belief functions can also be derived from models for dependent evidence.

Suppose, for example, that I do not judge Fred's testimony to be independent of the evidence provided by the thermometer. I exclude the possibility that Fred has tampered with the thermometer and also the possibility that there are factors affecting both Fred's truthfulness and the thermometer's accuracy. But suppose now that Fred does have regular access to the thermometer, and I think that he would likely know if it were not working. I know from experience that it just in situations like this, where something is awry, that Fred tends to let his fancy run free.

In this case, I would not assign the elements of S the probabilities given in the second column of Table 1. Instead, I might assign the probabilities given in the second column of Table 2. These probabilities follow from my judgment that Fred is truthful 80% of the time and that the thermometer has a 99%

chance of working, together with the further judgment that Fred has a 90% chance of being careless if the thermometer is not working.

When I apply (1) with the posterior probabilities given in Table 2, I obtain the degrees of belief

$$\mathrm{Bel}(\{yes\})=.005 \quad \text{and} \quad \mathrm{Bel}(\{no\})=.95.$$

These differ from (4), even though the belief functions based on the separate items of evidence will still be given by (2) and (5).

2.5. Conclusion. I would like to emphasize that nothing in the philosophy of constructive probability or the language of belief functions requires us to deny the fact that Baysian arguments are often valuable and convincing. The examples I have just discussed were designed to convince the reader that belief-function arguments are sometimes more convincing than Bayesian arguments, but I am not claiming that this is always or even usually the case. What the language of belief functions does require us to reject is the philosophy according to which use of the Bayesian language is normative.

s	Probability of s		Elements of T
	Initial	Posterior	compatible with s
(truthful,working)	.799	0	--
(truthful,not)	.001	.005	yes
(careless,working)	.191	.950	no
(careless,not)	.009	.045	yes,no

Table 2.

From a technical point of view, the language of belief functions is a generalization of the Bayesian language. But as our examples illustrate, the spirit of the language of belief functions can be distinguished from the spirit of the Bayesian language by saying that a belief-function argument involves a probability model for the evidence bearing on a question, while a Bayesian argument involves a probability model for the answer to the question.

Of course, the Bayesian language can also model evidence. As we have seen in our examples, the probability judgments made in a belief-function argument can usually be adapted to a Bayesian argument that models both the answer to the question and the evidence for it by assessing prior probabilities for the answer and conditional probabilities for the evidence given the answer. The only problem is that we may lack the evidence needed to make all the judgments required by this Bayesian argument convincing. Thus we may say that the advantage gained by the belief-function generalization of the Bayesian language is the ability to use certain kinds of incomplete probability models.

3. The Attempt to Use Probability in Production Systems

The field of expert systems developed within artficial intelligence from efforts to apply systems of production rules to practical problems. And the current interest in probability judgment in artificial intelligence began with efforts to incorporate probability judgments into production rules. In this section I review these efforts and relate them to what we learned

in the preceding section about the Bayesian and belief-function languages.

A production rule is simply an if-then statement, interpreted as an instruction for modifying the contents of a data base. When the rule is applied, the action specified by its right-hand side is taken if the condition on its left-hand side is found in the data base. A production system is a collection of production rules, which are repeatedly applied to the data base either in the same predetermined order or else in an order determined by some relatively simple principle. Production systems were used in programming languages in the early 1960's, and they were advanced as cognitive models by Newell and Simon in the late 1960's and early 1970's. (See, for example, Newell and Simon, 1965, and Newell, 1973.) Such systems are attractive as models for intelligence because their knowledge is represented in a modular way and is readily available for use. Each rule represents a discrete chunk of knowledge. Such a chunk can be added to or removed from the system without disrupting its ability to use the other chunks, and the system regularly checks all the chunks for their relevance to the problem at hand. (For a fuller account of production systems, see Davis and King, 1984.)

When artificial intelligence workers undertook, in the 1970's, to cast various bodies of practical knowledge in the form of production rules, they found that in many fields knowledge cannot be encoded in the form of unqualified if-then statements. Instead, probability statements seem to be required: "If E_1, $E_2, \ldots E_n$, then probably (or usually or almost certainly) H." So these workers found themselves trying to use production systems

to manipulate probability judgments.

Many tacks were taken in the effort to use probability in production systems, but I would like to emphasize two important lines of development. One of these begins with PROSPECTOR and leads to Pearl and Kim's elegant work on the propagation of Bayesian probability judgments in networks, while the other begins with the certainty factors of MYCIN and leads to the use of belief functions in hierarchical diagnosis. I will review these two lines of development in turn.

3.1. Bayesian Networks. The artificial intelligence workers at SRI who developed the PROSPECTOR system for geological exploration in the middle 1970's thought of production rules as a means for propagating probabilities through a network going from evidence to hypotheses. Figure 1, taken from Duda et al. (1976), gives an example of such a network; here E_i denotes an item of evidence, and H_i denotes a hypothesis. The idea is that the user of the system should specify that some of the E_i at the bottom of the network are true and some are false, or should make probability judgments about them, and the production rules, corresponding to conditional probabilities for the links in the network, should propagate these probability judgments through the network to produce judgments of the probabilities of the hypotheses.

The first thing the PROSPECTOR workers noticed about this scheme was that a Bayesian calculation of probabilities for the hypotheses would require more than conditional probabilities corresponding to the links and probabilities for the evidence nodes at the bottom; it would also require prior probabilities

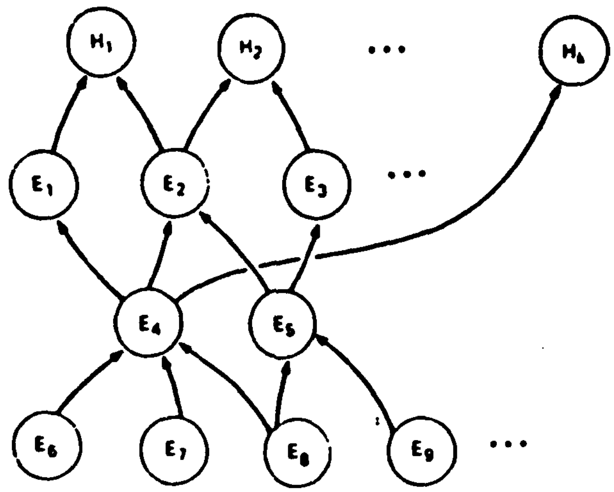


Figure 1.

for the hypotheses and the other evidence nodes. So they abandoned the idea of a pure production system at the outset by requiring that the expert knowledge in the system should also include these prior probabilities.

These workers retained from the production system picture, however, the idea that an expert's knowledge should come in discrete modular chunks. They wanted to be able to elicit from the expert statements of the form, "If E_i and E_j and ..., then E_r with probability p," without constraining the expert as to how these statements should fit together. This meant that they still faced problems in putting these chunks of knowledge together into a calculation of the probabilities of the hypotheses. Here are three of their problems: (1) The conditional probabilities elicited may not be sufficient to determine a joint probability

measure over all the E's and H's. If the expert is thinking in terms of the network in Figure 1, for example, he may give rules corresponding to $P(E_5|E_8)$ and $P(E_5|E_9)$ but neglect or feel unable to give a rule coresponding to $P(E_5|E_8 \& E_9)$. (2) The conditional probabilities that are given may be inconsistent. (3) The network may have cycles, which will cause trouble when propagation is attempted.

These problems were handled in PROSPECTOR in relatively ad hoc ways. Apparently problem (1) was handled partly by independence assumptions and partly by max-min rules reminiscent of the theory of fuzzy sets. Problem (2) was handled by formulating rules of propagation which did not always accord with Bayesian principles but which were insensitive to some kinds of inconsistencies. Problem (3) was handled by arbitrarily rejecting new production rules when they would introduce cycles into the network already constructed.

PROSPECTOR behaved in a reasonably intelligent way. But the ad hoc character of its procedures made many people ask whether a similar propagation of probabilities might be carried out in a more thoroughly Bayesian way. This question has been answered by Pearl (1982) and Kim (1983).

As Pearl and Kim show, we can make sense of the independence assumptions needed to construct a probability measure over a network from simple conditional probabilities and we can propagate updated probabilities through the network in a simple and elegant way provided that the network has a causal interpretation and a relatively simple form; it must be a simple directed tree or else a slightly more general directed graph called a Chow

tree. (Pearl (1982) treated the case of the simple tree, and Kim (1983) treated the case of the Chow tree.)

A Chow tree is simply a connected and directed graph such that there is no cycle in the corresponding unconnected graph; an example is shown in Figure 2. In Pearl and Kim's work,nodes of the tree correspond to random variables, and the directions of the links are interpreted as directions of causation. Thus each variable is influenced by the variables above it in the graph and influences the variables below it. An observation of the value of one variable is diagnostic evidence about the value of a higher variable and causal evidence about the value of a lower variable.

Once a Chow tree is constructed for a problem, the construction of a probability measure over it and the updating of the measure are straightforward. Given Kim's independence conditions, which are reasonable in the causal context, a measure over

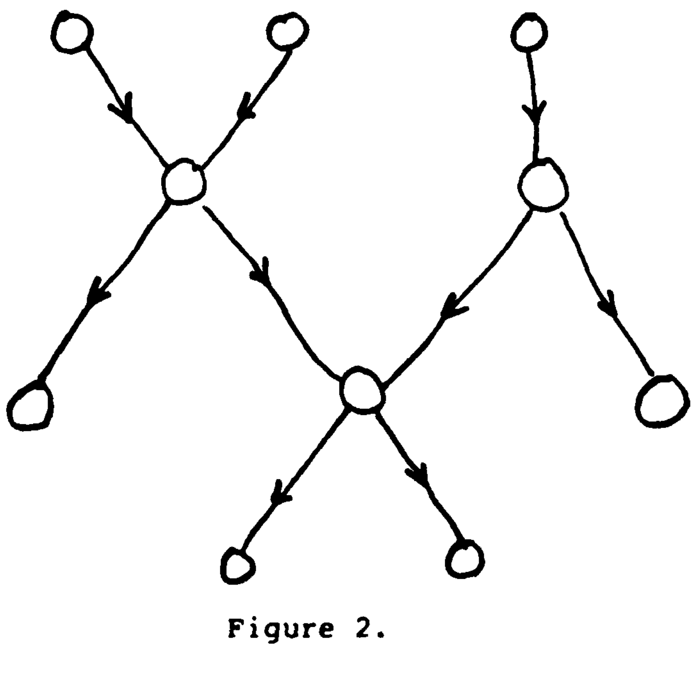


Figure 2.

the tree can be constructed from prior probabilities for the topmost nodes and conditional probabilities for all the links. Moreover, this construction is straightforward; there are no complicated consistency conditions that the conditional probabilities must meet. Once construction is completed, the measure can be stored and updated locally. At each node we store information about the conditional probabilities corresponding to incoming and outgoing links, the current probability measures for the variable at the node and the variables at neighboring higher nodes, and likelihood-type information from neighboring lower nodes. When the value of a variable is then observed, this information can be propagated through the network to update the entire probability measure in one pass. All computations are made locally, with each node communicating only updated local information to its neighbors.

An obvious shortcoming of this elegant scheme is its restriction to Chow trees. In few problems will the causal relations that we think important take so simple a form. Kim suggests that we might use such Chow trees as approximations to more realistic models; first elicit a probability measure on a more complicated graph from an expert, and then choose the Chow tree that best approximates this more complicated measure (Chow and Liu, 1968). This suggestion does not seem very satisfactory, however. We are given no reason to hope that the approximation will be satisfactory, and perhaps more importantly, the constructive nature of the initial probability measure is put into question. In a Chow tree the initial probability measure can be constructed from probability and conditional probability judg-

ments without concerns about consistency, but in a more general graph consistency conditions will be so complicated that it will be impossible for us to hope they will be met unless we pretend that we are indeed eliciting a measure instead of constructing one.

Another obvious shortcoming is the restriction to thoroughly causal models. Kim notes this problem as follows: "Although causal relationship is the most important one in situation assessment decision-making, it alone is insufficient to achieve an expert level of performance. Additional studies are needed to find ways of integrating causal relationships with other kinds of relationships to infer more valid conclusions."

In a sense, of course, all evidence is causal. We can always construct a model that relates the facts we observe to deeper causes and also relates these causes to the questions that interest us. The difficulty is that we may lack the evidence needed to make good probability judgments relative to such a model. The point that causal models are insufficient is really, therefore, subsidiary to the more general point, made in section 2 above, that we sometimes lack the evidence needed for a convincing Bayesian argument.

3.2. Certainty Factors and Belief Functions. Though I have begun my discussion of the effort to put probability in production systems with the PROSPECTOR story, the work on the MYCIN system for medical diagnosis began earlier and has been more extensive. The story of the MYCIN effort has been told in a recent book (Buchanan and Shortliffe, 1984), which includes ex-

tensive discussion of the certainty factors that were used by MYCIN and the relation of these certainty factors to belief functions.

MYCIN departed from the pure production system picture by using a backward-chaining strategy to select production rules to apply. This means that it selected rules by comparing their right-hand sides to goals instead of comparing their left-hand sides to statements already accepted. If the right-hand side of a rule matched a goal, its left-hand side was then established as a goal, so that there was a step-by-step process backwards from conclusions to the knowledge needed to establish them.

MYCIN also differed from PROSPECTOR in that the MYCIN workers rejected at the outset the idea that the numerical probability judgments associated with the rules could or should be understood in Bayesian terms. They emphasized this point by calling these numbers "certainty factors" rather than probabilities. And they formulated their own rules for combining these certainty factors.

In spirit, and to a considerable extent in form, these rules were quite like special cases of Dempster's rule for combining independent belief functions. I would explain this coincidence by saying that in developing their calculus for certainty factors, Shortliffe and Buchanan were trying to model the probabilistic nature of evidence while avoiding the complete probability models needed for Bayesian arguments.

In recent work (Gordon and Shortliffe, 1984, 1985), the MYCIN workers have taken a close look at the similarity between

the calculus of certainty factors and the language of belief functions and have asked how belief functions can contribute further to the MYCIN project. They have drawn two main conclusions. First, it is sensible to modify some of the rules for certainty factors to put these rules into more exact agreement with the rules for belief functions. Second, the diagnosis problem that was central to MYCIN can be understood more clearly in terms of belief functions if it is explicitly expressed as a problem involving hierarchical hypotheses.

The term "hierarchical hypotheses" refers to the fact that the items of evidence in a diagnostic problem tend to support directly only certain subsets of the frame of discernment, subsets which can be arranged in a tree. Figure 3, taken from Gordon and Shortliffe (1984), illustrates the point. The four nodes at the bottom of this tree represent four distinct causes of cholestatic jaundice; they form the frame of discernment for the diagnostic problem. Some items of evidence may directly

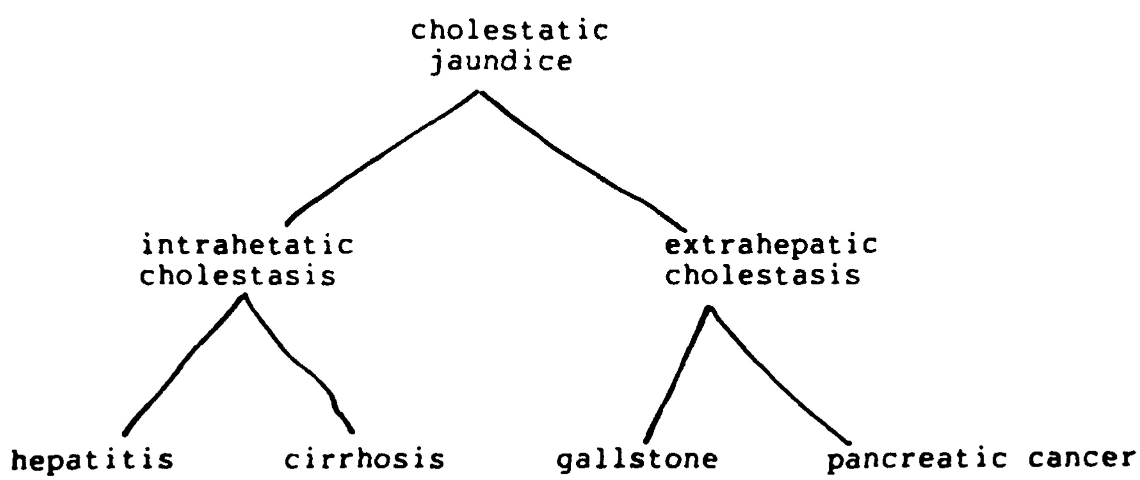


Figure 3.

support (or directly refute) one of these causes for a particular patient's juandice. Other evidence may be less specific. There may, for example, be evidence that the jaundice is due to an intrinsic liver problem, either hepatitis or cirrhosis. On the other hand, it is hard to imagine a single item of medical evidence supporting the subset {cirrhosis, gallstone} without supporting one of these more directly; this is reflected by the fact that this subset does not correspond to an intermediate node of the tree.

This picture suggests that a belief-function argument based on such medical evidence may involve combining many belief functions by Dempster's rule, where each belief function is a simple support function focused on a subset in the tree or its complement. (A simple support function is a belief functon obtained from (1) when S has only two elements and one of these is compatible with all the elements of T.)

Though the tree structure provides a conceptual simplification of the problem, the combination of simple support functions corresponding to subsets in the tree and their complements can result in a very complex belief function and hence threatens to involve prohibitive computation. Gordon and Shortliffe (1985) have proposed a modification of Dempster's rule for this situation that would involve less computation and would often give similar results.

I believe that this modification can be avoided. Though full computation of the belief function resulting from Dempster's rule often would be prohibitive, we would seldom need full computation. Usually we would need only to identify subsets in the

genuine intelligence to their designers and users. Their designers will have to design the forms of probability argument for the particular problem, and their users will have to supply the probability judgments.

4. The Construction of Arguments

I have emphasized that a genuine capacity for probability judgment in an artificial intelligence would involve both the ability to generate numerical probability judgments and the ability to design probability arguments. How might these abilities be programmed? We do not have an answer, but we should start thinking about the question.

As the result of the work by psychologists during the past decade, especially the work of Kahneman and Tversky (see Kahneman, Slovic, and Tversky, 1982), we do have some ideas about how people generate numerical probability judgments. They conduct internal sampling experiments, they make similarity judgments, they construct causal models and perform mental simulations with these models, they consider typical values and discount or adjust these, and so on. An obvious and appropriate strategy for artificial intelligence is to try to implement these heuristics.

The heuristics sometimes lead to systematic mistakes or biases, and it is by demonstrating these biases that the psychologists have convinced us that people use them. There is a tendency, therefore, to think that people are doing something suboptimal or unnormative when they use them. Indeed, proponents of the Bayesian philosophy frequently assert that the psychological work only demonstrates what people do do and is irrelevant to

what people should do. Presumably this means that instead of using the heuristics, they should first realize that it is normative for them to have preferences satisfying Savage's postulates, then decide to pretend that they do have such preferences, and then try to figure out what they are.

When we face up to the artificial intelligence problem, however, we see that the heuristics are really all we have. People have to use such heuristics if they are to make quick probability judgments about questions they have not previously considered, and our programs will also have to use them if they are going to be equally flexible. The challenge is to figure out how to use the heuristics well enough that using them will not usually cause mistakes.

Implementing the heuristics involves us, of course, in all the issues of knowledge representation, for we must have a flexible way of matching the problem about which we want to make a judgment with similar problems in our memory.

It is more difficult to say anything about how we might build the ability to design probability judgments. The lesson from section 3 is clear, though: the chunks that we try to fit together when we search for a convincing argument must be larger than the chunks represented by probabilistic production rules. It is also clear that the ability to construct convincing probability arguments must include an ability to evaluate whether a probability argument is convincing.

Though these questions are difficult, they should be taken as a challenge by students of probability. I believe that proba-

tree that have high degrees of belief and to compute these degrees of belief. This should usually be achievable by careful computational strategies.

Violations of the independence assumptions needed for Dempster's rule may pose a more important problem. It seems unlikely that the uncertainties involved in a very large number of items of medical evidence will all be independent. This does not mean that a belief-function analysis will be impossible or unsatisfactory, but it does mean that a satisfactory belief-function analysis may require modelling dependencies in the evidence.

The two needs identified here, the need for effective computational strategies and the need for models for dependent evidence, also arise in many other contexts.

3.3. Conclusion. The preceding look at attempts to use probability judgment in expert systems justifies, I think, a general conclusion: probability judgment in expert systems is very much like probability judgment everywhere else. Though the builders of MYCIN and PROSPECTOR worked in the context of artificial intelligence ideas, the thinking about probability judgment that has emerged from this work ten years later does not have a distinctive artificial intelligence flavor.

The general issues about probability judgment that we identified in section 2 above all re-appear in the expert systems work. In expert systems, as elsewhere, probability judgment is constructive and requires an overall design. It is generally possible to provide such a design within the Bayesian language, but Bayesian designs often demand judgments for which we do not

have adequate evidence. And belief-function analyses often require models for dependent evidence.

Production systems were attractive to the artificial intelligence community because these systems seemed to have the flexibility in acquiring and using knowledge that seems characteristic of intelligence. But it seems fair to say that the attempt to incorporate probability judgment into production systems failed. PROSPECTOR and MYCIN themselves retained a good deal of the flavor of production systems, but little of this is left in the work of Pearl and Kim or in the proposal to use belief functions in hierarchical trees. It appears that probability judgment simply does not have the modular character that made production systems so attractive. Almost always, probability judgment involves not only individual numerical judgments but also judgments about how these can be put together. This is because probability judgment consists, in the final analysis, of a comparison of an actual problem to a scale of canonical examples.

(Expert systems that are based on production rules without probability judgment have been very successful; R1 and DART are often cited as examples of commercial success. The developers of these systems have been heard to say, however, that their methodology is best used in problems where probability judgment is not needed.)

I would suggest that the expert systems we see using probability in the near future are not likely to have the flexibility and judgmental capacity that we associate with genuine intelligence. Instead, these systems will continue to leave the work of

bility judgment will turn out to be possible and important in artificial intelligence, but the extent of its ultimate usefulness cannot be taken for granted; it must be demonstrated.

References

Barr, A., and Feigenbaum, E. A. (1981) The Handbook of Artificial Intelligence, Volumes I and II. William Kaufmann, Inc., Los Altos, California.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, Series A 143:383-430.

Buchanan, B. G., and Shortliffe, E. H. (1984) Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Reading, Massachusetts.

Chow, C. K., and Liu, C. N. (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory IT-14:462-467.

Cohen, P. R., and Feigenbaum, E. A. (1982) The Handbook of Artificial Intelligence, Volume III. William Kaufmann, Inc., Los Altos, California.

Cohen, P. R. (1983) Heuristic reasoning about uncertainty: an artificial intelligence approach. Report No. STAN-CS-83-986, Department of Computer Science, Stanford University.

Davis, R., and King, J. J. (1984) The origin of rule-based systems in AI. In Buchanan and Shortliffe, pp. 20-52.

Doyle, J. (1979) A truth maintenance system. Artificial Intelligence 16:257-294.

Duda, R. O., et al. (1976) Subjective Bayesian methods for rule-based inference systems. In _AFIPS Conference Proceedings of the 1976 National Computer Conference_, Volume 45 (New York), pp. 1075-1082.

Gordon, J., and Shortliffe, E. H. (1984) The Dempster-Shafer theory of evidence. In Buchanan and Shortliffe, pp. 272-292.

Gordon, J., and Shortliffe, E. H. (1985) A method for managing evidential reasoning in hierarchical hypothesis spaces. To appear in _Artificial Intelligence_.

Kahneman, D., Slovic, P., and Tversky, A. (1979) _Judgments under Uncertainty: Heuristics and Biases_. Cambridge University Press.

Kim, J. H. (1983) _CONVINCE: A Conversational Inference Consolidation Engine_. Doctoral dissertation, Computer Science, University of California at Los Angeles.

Marr, D. (1982) _Vision_. Freeman, San Francisco.

Newell, A., (1973) Production systems: models of control structures. In _Visual Information Processing_, W. G. Chase, ed., pp. 463-526. Academic Press, New York.

Newell, A., and Simon, H. A. (1965) An example of human chess play in the light of chess playing programs. In _Progress in Biocybernetics_, N. Wiener and J. P. Schade, eds. Elsevier, Amsterdam.

Pearl, J. (1982) Distributed Bayesian belief maintenance. _Proceedings of the Second National Conference on Artificial Intelligence_. William Kaufmann, Inc., Los Altos, Califor-

nia.

Shafer, G. (1976) A Mathematical Theory of Evidence. Princeton University Press.

Shafer, G. (1981) Constructive probability. Synthese 48:1-60.

Shafer, G. (1985) Belief functions and possibility measures. To appear in The Analysis of Fuzzy Information, Volume 1, J. C. Bezdek, ed., CRC Press.

Shafer, G., and Tversky, A. (1985) Languages and designs for probability judgment. To appear in Cognitive Science.

TRANSCRIPT OF ORAL PRESENTATION BY GLENN SHAFER: PROBABILITY JUDGMENT IN ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEMS

DR. SHAFER: I would like to speak a little more broadly than the topic of belief functions.

I think belief function argument is a kind of probability argument. To understand what its role is, or could be, in artificial intelligence and expert systems I'd like to talk a little more generally also about the role of probability in artificial intelligence.

Here's a list of topics I would like to discuss. I'd like to talk about what belief function arguments are, and contrast them with the Bayesian arguments. A belief function argument, as I said, is a probability argument but it is often distinct from the Bayesian argument. I would like to talk about the general philosophy of what I call constructive probability, according to which a probability judgment always involves a comparison of an actual problem to a scale of canonical examples, where the canonical examples are usually familiar examples from the picture of games of chance.

I'd like to talk about what the role of probability is in expert systems. I think it often is essential, but expert systems seem to require flexibility which we are not accustomed to with probability arguments.

Finally, I would like to talk about the role of probability in artificial intelligence proper, as distinguished from expert systems, which I think is an area which has not been explored very much in the course of the recent interest in the subject.

In fact, I think it would be interesting to start with the question "why has there been so little probability in artificial intelligence?"

Lately we've seen an explosion of interest in the topic but that has to be contrasted with the previous 20 years when there was practically none on the part of the artificial intelligence community.

Why was that? Well, I think there are some basic historical reasons. First of all, artificial intelligence began in the 1950s, very self-consciously, by contrasting itself with what computers could then do, which was crunch numbers.

The idea was that intelligence surely involves something more than that, some kind of more general symbolic manipulation, and so numbers and probabilities in particular were something that they were not, by definition, interested in.

That attitude was buttressed by kind of a folk argument, something that you don't encounter in print, at least very often, but you often encounter in talking to people in the artificial intelligence community. I call it the input-output argument, which says that artificial intelligence is an attempt to imitate human intelligence and in the case of human intelligence the inputs are nonnumerical and the outputs are nonnumerical.

What use is there for numbers in between? Perhaps you could phrase something that goes on in between in terms of numbers, but why bother?

Basically the intelligence, as it were, must be a program of some kind for going from inputs to outputs. Why not just tell what that program really is doing? What it's doing must surely have nothing to do with numbers, so why talk about probabilities at all in artificial intelligence?

I think that argument is further buttressed by some developments within artificial intelligence which present themselves in some sense as alternatives to probability. One of those is the idea of nonmonotonic logic, a different way of handling uncertainty.

More recently a fellow named Paul Cohen who is now at U Mass Amherst has talked about following more closely the input-output argument and talked about giving explicitly the reasons or the endorsements that a program would need to take certain actions -- again a self-conscious alternative to probability.

Why the current interest in probability? Why the change?

Well, it is pretty clear that the cause of this interest is the expert system idea.

The artificial intelligence people got interested in using their ideas about knowledge representation to build programs which they would call intelligence systems that would make an expert's knowledge available to a nonexpert user, and as soon as they started looking in that direction they found that in some domains expert knowledge seems to have to involve probability judgments. You can't make absolute what the expert knows or is able to tell you. It doesn't seem to be in an absolute form, if such-and-such then such-and-such. It's only if such-and-such, then probably such-and-such.

Since you are going here on a human judgment, you can put human inputs, and therefore numerical inputs, into the system. Thus we've gotten away from the assumption, in expert systems as opposed to the original artificial intelligence problem, that the inputs are nonnumerical, and immediately have both a need and an opening for probability.

As I said at the beginning, I think of belief function arguments as a special kind of probability argument. Why have the AI people been so interested in belief function arguments?

I once visited a computer science department about a year ago and they knew I was a statistician so they took me upstairs and introduced me to the chairman of the statistics department. After a little bit of gossip he turned to me very gravely and said, "Why are these people so interested in what you do?"

This is the question I want to ask here -- why are these people so interested in belief functions?

I think the answer has to do with their hope for a modular representation of probabilistic knowledge.

At the time that expert systems were getting started in the 1970s, a fundamental idea in the artificial intelligence community (I haven't been part of that community and I think there are some people here that are and they might be able to correct me) was that the flexibility of human intelligence might be explained in terms of certain modularity in the representation of knowledge. Apparently things are arranged in our heads in such a way that individual items of knowledge can be added or removed without disrupting the whole system, which contrasts very strongly with what we are accustomed to in terms of structured computer programming. You can't take one line out of a FORTRAN program and expect the thing to work.

It also contrasts very strongly with the Bayesian probability argument. You can't take one of the inputs out of a Bayesian argument and expect the thing to work.

Somehow human intelligence does go that way. The structures are built in such a way that you can get along without any particular thing. With the talk about combination of evidence, discrete items of evidence being put together, and being able to make judgments on the basis of limited evidence, belief functions seem to offer some hope in that direction.

That's my perception of why this interest has come up. I don't know how far I'll get this morning, but I'd like to offer you part of my conclusion in advance.

The conclusion is (a biased one, of course) that I think belief functions are more flexible and more modular in some respects than Bayesian arguments, but I don't think they are as modular as what the artificial intelligence people were looking for in the 1970s.

I think that all probability argument involves an overall design, in some sense. We can't get the kind of modularity that the AI community was looking for when they were building expert systems based on production rules. When we really get into probability, expert systems are going to look different than they do now. There are a lot of expert systems and only some of them came from the artificial

intelligence community. The ones that came from the artificial intelligence community, like DART and XCON, are largely based on the idea of production rules and do have this modularity which I don't think we're going to have when we get successful systems based on probability.

Another conclusion is that in the case of probability, the expert systems we see in the future may not be that strongly influenced by ideas coming out of the artificial intelligence community at the present time.

With that beginning, I'd like to go to the part of the talk that I was asked to give, which is about belief functions.

To talk about belief functions quickly, I would like to first introduce the idea of a frame of discernment, which is a list of possible answers to a question. It's what we are accustomed to calling a sample space in statistics, though sometimes it's a parameter space, any space that we could put a probability measure on.

I want to talk about the idea of a compatibility relation between two frames. Say S is one frame and T is another and we have an element little s in S and a little t in T. Let's write sCt to mean that little s is compatible with little t. That means that little s could be the answer to the first question and at the same time little t could be the answer to the second question. It's possible for s and t both to be the answers.

Let me quickly in an abstract way tell you what a belief function is, in case you aren't clued in.

The idea is that you start with a probability measure P (perhaps based on frequencies from your experience) on S, and then you get the belief function on the second frame T, by setting $Bel(B) = P\{s\ sCt \text{ implies } t \in B\}$.

The statement here is any probability on S that can only be associated with answers to the second question which are in B should count as a reason to believe that the answer to the second question is in B.

I didn't say that very smoothly. Let me see if I have it written down any better.

Bel (B) measures the reason we have to believe B based on the probability in the compatibility relation between S and T.

Well, that's only if we didn't know anything else. The idea is this is the reason we have based on P and its compatibility relation, so the idea is we might have some other evidence that more directly bears on T but if we want to leave that aside for the moment or if we don't have it, or if we just want to talk about what support we have for B based on the evidence summarized by P and by the compatibility relation between S and T, if we just want to make judgments based on that evidence, this seems to be the kind of judgment we want to make.

Let me give an example. Fred comes sauntering over to my desk and he's about to tell me something and the question in my mind is is this going to be on the level or is it just, like he does sometimes, talking without paying attention. My experience is that he's truthful 80 percent of the time and he's careless 20 percent of the time and I generally can't tell what he's up to.

The second question that's on my mind is whether the streets are icy outside, as sometimes happens in Kansas. The possible answers to that question are yes and no, so we have two questions. Here are the possible answers to them and I have a probability measure on the first. I might have some evidence more directly about whether the streets are icy but I want to think about the situation where my evidence is just Fred's telling me they are icy. (see Slide 1)

Well, since I have this high confidence in what Fred says, about 80 percent, that seems to give me some reason just in itself to belief the streets are icy.

Now what's the compatibility relation? Well, once Fred has told me this, then Fred's being truthful is compatible only with the streets being icy, but Fred's being careless is compatible with either one. That's what this says here. That's our compatibility relation. (see slide 2)

Let's apply that formula I wrote down to this situation. Here's the formula. In this case we have probability .8 on truthful, .2 on careless, truthful is only compatible with yes, but careless is compatible with yes and no.

What does this formula tell us? Well, in this case, B just consists of a single point, "yes," so here I just want all s's that are compatible only with yes and that's this one, so my degree of belief for yes is eight-tenths.

On the other hand, there are no s's that are compatible only with "no" because careless is compatible with both, so I don't have anything here. My degree of belief for no is zero.

So on the basis of Fred's testimony alone, I have an eight-tenths reason to believe that the streets are icy outside and zero reason not to believe it. That's the basic idea of belief functions. (see slide 3)

In some sense a belief function argument is just a probability argument. What's going on is that we're putting the probabilities on a space or a frame in the background and we are looking at the implications for a different frame which more directly interests us. So in a sense, all the usual things you can do with probability you can also do with belief functions. It's just that the probability stuff you're doing back here and it's after you get that done that you then look at what's going on in the frame of interest.

This also makes it clear that belief functions are a generalization of the usual Bayesian approach to probability because you could after all consider the special case where S and T are the same. Then you're working directly. The compatibility relation in that case would just be that an element here is compatible only with itself in the second copy so --

DR. SINGPURWALLA: Why is truthful not compatible with no?

DR. SHAFER: Because of what he said. This is after the fact. He said it's icy outside. It's his statement. Our knowledge establishes the compatibility relation between the two frames.

DR. DeGROOT: Since there is a break, let me just interrupt for one moment and say something I should have said before we started, and that is I think during the talk you should feel free to interrupt but only for clarification questions. You should be able to do -- if that's okay with you I think the audience should be able to do that but we'll give you the traditional speaker's privilege of refusing to answer questions that you don't like during the talk and after the talk you no longer have that privilege of refusing to answer.

DR. SHAFER: Let us contrast what I was just doing with a Bayesian analysis.

There are a lot of the Bayesian analyses for a given problem but what would you do if you wanted to take this eight-tenths and two-tenths that I was talking about and extend it to a full Bayesian argument?

Well, you would need two things. First of all, the point that is most often emphasized when we're contrasting Bayesian with traditional statistical arguments, you need prior probabilities. What is your other evidence for whether or not it's icy outside? Let's say you have probability p without Fred's testimony, probability p that it is icy, probability 1-p that it isn't icy.

There's another thing you need too in order to do the Bayesian argument. You need to break down this "careless" into two cases, where he's careless but what he says is true, and where he's careless and what he says is false. You have to put that in your probability model too. So you put q of that two-tenths into careless but true, and 1-q in careless but false. Then you can do the Bayesian thing.

That can be represented by Bayes' theorem but let's not. Let's say that if you put these judgments together before you heard what Fred said, S and T were independent, so you could form the product measure on S and T. Then you would condition on the knowledge that you got from what Fred said. (see slide 4)

If they're independent then you multiply the 8/10 times the p to get the probability that Fred is going to be truthful and also it's icy outside, et cetera, these numbers here represent the product measure, and then of course you want to condition on the fact that Fred does say that it is icy, in which case he can't be truthful at the same time as

it's not icy and also he can't be saying something false at the same time it's not icy, et cetera, these three things are crossed out and you renormalize as usual with conditioning and so this would be your probability for "yes, it is icy outside."

Now this gives a different answer in general than the belief function argument for your degree of belief that it's icy outside. What it actually would be depends on what p and q are.

The belief function argument, contrasted with this, uses a less complete probability model.

Now, of course, if this probability mode were somehow correct, if things were really developing by chance and these really were the chances, of course this answer would be right. Clearly, the belief function answer would be wrong.

There's a tendency to say, well, maybe we don't know what p and q are, but they must be something, and whatever they are it isn't likely this formula will agree with the belief function answer, so the belief function answer must be wrong.

My answer to that argument is that basically this chance thing is not in any sense what's really going on. What we're doing is we're modeling what's really going on by comparison to the picture of chance. I may refuse to go all the way and do the Bayesian analysis. The question is whether or not I feel I have the evidence to support these judgments.

The belief function idea is that perhaps in some situations we can gain more flexibility by making only comparisons to the picture of chance for which we have enough evidence. Some other comparisons we feel we don't have any evidence for. Perhaps there is some regularity to Fred's behavior in terms of whether he's being careless or truthful but there may not be any regularity that we can make out as to whether or not what he says is true when he's being careless and we don't want to model that by comparison to a chance picture.

It's not as if there's some comparison which is right, or as if there are some chances which are right and we just don't know them. Rather, we don't want to make the comparison to the chance picture at all.

With that attitude, we don't want to make a comparison we don't feel is a good comparison. With that attitude this number here has no reality. It's not like there are some p's and q's we don't know. There just aren't any. If there were and somebody else knew what they were, we certainly wouldn't want to bet with that person if we were doing the belief function analysis, but then of course we don't want to bet with people that know more than we do in general.

VOICE: Is it in order to ask what you mean by "chance"?

DR. SHAFER: Yeah. There are these balls and urns and dice and things in physics. I'd better hurry through a couple of other examples.

I want to talk about combining evidence because this is essential to the belief function picture.

Let's say we did have some other evidence about whether it's icy or not that we want to bring in. Fred's testimony is one item of evidence. Let's say I had another. Suppose I have an indoor-outdoor thermometer near my desk and the indoor-outdoor thermometer says it's 31 degrees and I know from experience that you can't get ice on the streets when it's 31 degrees with what traffic we have going by.

My thermometer is very reliable, those things are pretty reliable devices, and this one has been pretty well calibrated for a long time. I have a lot more confidence in it than I do in Fred and it's an argument for it not being icy outside.

Let's say we wanted to put those two arguments together. Well, again, I want to have one space S that I do my probability calculations on, keeping it distinct from the space T corresponding to the question I'm interested in, whether or not it's icy.

Here I'm just working with the question as to whether or not Fred has been truthful or careless and whether or not that thermometer is working. (see Slide 5)

I have a 99 percent probability that the thermometer is working, one percent probability that maybe something is wrong with it.

What do I do? Well, again I make a judgment back here in the probability area. I make a judgment of independence, so I construct a probability measure there. Eight times .99 is .792 probability that Fred is going to be truthful and the thermometer is working.

This is the probability I would construct before I hear what Fred said and see what the thermometer says. After I hear what Fred says and see what the thermometer says I have to make a change because I know that Fred and the thermometer are contradicting each other. This possibility can't have happened so I eliminate it; I condition on eliminating it and renormalize just as in the Bayesian story, except here I'm working in the background and not directly on the frame of interest.

These are the four possibilities in the background. These are the elements of T that are compatible with them after I get the knowledge of what Fred said and what the thermometer said. There's nothing that's compatible with the thermometer working and Fred being truthful, because they contradicted each other. (see slide 6)

If Fred is truthful and the thermometer is not working, then that means it is icy outside. If Fred is being careless and the thermometer is working, that means it's not icy outside because the thermometer says it isn't.

If Fred is being careless and the thermometer is not working, I don't know what's going on then. That is compatible both with it being icy and with it not being icy.

These are the initial probabilities based on that product measure. I condition by eliminating this one and renormalizing. I get .95 here and .04 here and .01 here so my degree of belief on yes is the .04 and my degree of belief on no is the .95.

The point is I had more confidence in the thermometer than I did in Fred, so in spite of Fred's testimony I have a degree of belief of .95 in it's working.

This is an example of Dempster's rule for combining belief functions in general. The general story is you form in the background a product measure, and you condition it on what's implied by the compatibility relation between the two frames.

One more example: Let me talk about dependent evidence. In the preceding example we had independent evidence in the background, so we formed a product measure. But in many problems we don't have that kind of independence. Let me just tell a story where we don't see what would then happen with the belief functions.

The story is basically the same. I make a judgment that Fred is 80 percent reliable, and the thermometer is 99 percent reliable. But this time I do not make the judgment of independence.

In the first story, I was saying, well, maybe Fred doesn't even see the thermometer. Fred doesn't have anything to do with the thermometer. It's on my desk.

In this story, let's say there is some dependence. Maybe Fred is the one that pays more attention to the thermometer that I do, and in fact if it weren't working it's likely that he would have known about. He probably would have known yesterday that it wasn't working, and it's precisely in cases like this that Fred lets his fancy run loose.

If the thermometer had been working, it's not so likely that he would have gone mumbling around about whether it was icy or not but with the thermometer not working he feels he sort of lets his own mind run free about what's going on. (see slide 7)

Let's say that there's a 90 percent chance of Fred being unreliable if the thermometer is not working. These three judgments are enough to construct a probability distribution. We have a marginal probability for his reliability and a conditional probability for his reliability and a marginal probability for the thermometer's reliability, and these are enough to determine a probability

distribution on on our space of four elements here and here are the probabilities that they determine.

Well, once again, after we see what Fred says and see what the thermometer says, this compatibility relation eliminates this possibility and so we renormalize and get the answers here, which are slightly different from the ones we had before.

The .95 is the same. These are the first ones. These are the second ones. The .95 are the same but whereas in the first case we had a .04 here now we have .005. (see Slide 8)

In this particular case, there is really no substantial difference in what the probability judgments are but conceptually there is quite a difference in what's going on because in the first case we have two items of independent evidence and in the second case we had two items of evidence which are dependent.

The idea is, as I said once already, we can do anything -- all the usual ideas of probability including independence or dependence are there for us to work with. It's just that they're going on in the background on a different frame instead of the one that's directly of interest.

So what's going on here? Well, as I say, we have these two frames, S and T. Now if I wanted to contrast the Bayesian with the belief function approaches I would say that in general the Bayesian argument acts as if the answer to the question of interest was determined by chance whereas the belief function argument acts as if the answer to a related question were determined by chance.

A constructive attitude is that we want to be explicit about the fact that what we're really doing (when we make a probability judgment of either kind) is comparing our actual situation to the picture of chance.

More precisely, we're fitting the actual situation to a scale of canonical examples, a scale of different pictures involving chance, where the numbers are different. The difference between the Bayesian and belief function stories is that we're using different scales of canonical examples to which we're comparing our actual situation.

In either case, there is in general a problem of design. Since we are comparing our whole problem to a picture of chance, we have to have a general design for how we're doing that.

In all these examples I've been giving, we didn't just take individual judgments and put them together in an arbitrary way. There was a general overall comparison, a general design.

I'm trying to say more things that there is room for in one talk but let me quickly shift back to the artificial intelligence story and try to give a little more depth to the comments I was making earlier about production rules.

Work in expert systems in artificial intelligence in the early '70s and on into the late '70s, as I mentioned before, was largely based on the idea of production rules.

What's a production rule? It's just an if-then statement. Why was it interesting to people in artificial intelligence?

Well, it was interesting because it does in fact offer a way to do very unstructured programming.

The idea is that perhaps you could represent your knowledge by a large set of if-then rules -- like if there is smoke then there is fire -- and perhaps you could apply those rules to current knowledge in a very unstructured way.

The idea is you put your current knowledge in what you would call a data base. So, for instance, in your data base might be the current knowledge that there is smoke.

Now here's a very unstructured way the system might operate. Suppose that what you do is you just start at the beginning and go through all your production rules and with every production rule you check whether the hypothesis matches something in the data base.

If it does, then you take the conclusion and put it in the data base, too, so when you come to this production rule you say, aha, there is smoke and you add to the data base the statement that there is fire, so then you go through the whole set of production rules and then you come back and start over.

Since there are new things in the data base, you may be able to do new things the second time around that you didn't do the first time around and if you have a lot of production rules in fact you can construct very complicated arguments in this way.

The nice thing about it is that you can do that without really worrying about the details of how you're going to structure it. You can just put enough production rules in there that the thing is going to work and in fact it might work even if you took a few out. Some of the production rules might only be helping you do things more quickly and not be essential for getting them done.

A lot of the expert systems that were built in artificial intelligence used this kind of a setup.

Of course the people that were working with this found themselves trying to apply it to domains where they couldn't really put in these absolute production rules, where all they could say were things like "probably." So the question began -- "Instead of just being able to say if there's smoke there's fire, we can only say if there's smoke then there is probably fire." Maybe putting some degree of probability here. How can you adapt the production rule system to that? So you say there's a probability there is smoke, and a probability on the production rule; then when you put "there is fire" in there you should have some probability connected with it, too.

So as you go through, you should be not only drawing conclusions, but you should also be propagating probabilities.

Well, the people that were doing this were not. They knew something about traditional ideas of probability but they were dealing with a problem that wasn't familiar, so they asked how can we do something that's sort of acceptable but follows this idea of propagating probabilities.

One well known example is the PROSPECTOR system. This was developed in the Stanford Research Institute. It was a system for geological exploration, and they were encoding geologists' expert knowledge with probability judgments attached with them, and they drew pictures like this. (see slide). If there's a certain kind of rock, then there is likely to be something else, et cetera. Certain items of evidence suggest other items of evidence. They drew pictures of nets where you go through items of evidence and eventually you get to the hypotheses that interest you. (see Slide 9)

You see the links would correspond to if - then statements that have probabilities attached to them. The links are the production rules. For people that are accustomed to thinking about probability the links seem to correspond to some sort of conditional probabilities. So how are you going to use these links to get from judgments down here to the judgments up here that you're interested in?

Well, the first thing that they noticed, of course, is that Bayesian analysis requires more just than conditional probabilities. You also have prior probabilities.

Well, okay, then you can have your expert give you not only conditional probabilities but also prior probabilities on all these nodes and then maybe your user could supply either the fact that certain of these items at the bottom were true, or perhaps some new probability judgment that they were true based on the specific case.

You want to take those new, either certainties or probabilities at the bottom, and propagate them through the system.

Well, this doesn't fit very well with anything we're accustomed to doing in probability theory, and this came out in terms of some problems that were perceived with what was going on.

One problem can be described as saying the conditional probabilities given by these links are not sufficient to determine - even with the prior probabilities, they may not be sufficient to determine the probability distribution on the whole space. They may not be sufficient because you have that conditional probability of this, given this, and this given this, and of this given this, but these are only pair-wise probability judgments. You don't have anything that involves three terms so that you can get the joint probabilities for larger groups of elements.

On the other hand, what you do have may well be inconsistent. If people just start throwing out these conditional probability judgments you may not have anything that is consistent with an overall probability distribution.

A third problem is that though there aren't any cycles in this picture, if you sit down and just take in the information that a geologist is willing to provide, you might have some conditional probabilities going that way and also some going this way and if you decide you are propagating from the bottom up you might find out that in the course of your propagation you're going around in circles, which doesn't make any sense for what you're trying to do.

The PROSPECTOR people dealt with these problems in various ways that were sort of ad hoc. I mean obviously you can deal with this insufficiency problem by making various kinds of independence assumptions. They did that.

From what they've published, or at least from what I've seen, it's not clear, certainly all the details are not clear to me, to some extent I think they used independence assumptions here. They also seem to have used some max/min types of rules.

The consistency problem, well, they solved that again, sort of, by making up their own kind of propagation rules, which really couldn't be squared completely with the usual probability arguments, but which did get away from some of the consistency problems.

As to the cycles, of course, they just put some rules in their system that said if the geologist volunteered some information that was going to create a cycle the system refuses to accept it or, if it does accept it, then it rejects -- takes something out that was already there, so you deal with that just by brute force.

Well, that was interesting enough to people in artificial intelligence that a lot of people asked the question is there some way you can do this better, is there some way you can do an honest Bayesian job with this propagation of probabilities business? I think that question has been answered pretty well by Judea Pearl of UCLA and his student Kim in some work they published in just the last couple of years

on the propagation of probabilities. They've settled, I think, just what kinds of assumptions are needed for this kind of propagation.

One assumption is that you can't have a cycle. To do it in a consistent Bayesian way, you have to have what is called a Chow tree, which is very nearly a tree. Your graph cannot have any cycles, not even any directed cycles.

Here you do have a cycle in a sense but you can't follow the cycle the way the arrows are going. So that was okay, but in a Chow tree you can't even have a cycle like this. Even this cycle was not allowed. A cycle like this was not allowed, even one that you can't follow around if you follow the arrows, so it's a pretty restricted system. (see Slide 10)

But if you do assume this kind of a Chow tree and you do put individual conditional probability judgments here, then you can actually make sense.

A second point, if you interpret this tree in a causal way so that you're thinking of the things down here as being causes of the things up here, then you can make sense of the conditional of the independence assumptions that are required in order to go from the pair-wise judgments to a complete probability distribution.

Also, you don't have any consistency problem so no matter what those individual judgments are, they're consistent and they do determine a probability distribution.

Third, Pearl and Kim have done this very elegantly, you can propagate the probabilities. These nodes correspond to random variables. So the arrows correspond not just to a single conditional probability but to a whole set of conditional probabilities, conditional probabilities for this variable given this one. So if you find out the value of any variable you can condition the whole system on that, and do so by local computations in one pass through the tree. You only have to store locally information about the neighbors below, what probability they're telling you, and the neighbors above, what likelihood information they're giving you, and that can all be done locally.

I think to most people's minds they have shown what can be done, rigorously and Bayesianly, in the direction that the PROSPECTOR people were trying to go.

VOICE: When you have a fairly complicated production system, you don't even know which way the tree can go, right? I mean beforehand.

MR. SHAFER: That's right. I think that's the question we come to here. Pearl and Kim showed what you can do in the direction that the PROSPECTOR people were trying to go, but I think the question there is what's left of the production system idea that you started out with, and I think the answer is not much.

You have to have a very structured picture and you don't have that modularity of knowledge that you did have. You don't have the flexibility and representation of knowledge.

The only flexibility you can get is in the flexibility your system might have in constructing trees like this by interacting with the user, the expert or nonexpert user.

Is there any distinctive AI flavor left in this? Well, I don't think so, not distinctive in the sense that this picture (that is, Pearl's pictures) is exploiting work done by Chow in operations research, and has gotten away from what was distinctively artificial intelligence.

Well, let me quickly look at another case of what came out of this expert system work in artificial intelligence in the '70s. The MYCIN program, which I'm sure most of you have heard about, was actually undertaken earlier than the PROSPECTOR program.

I talk about it second instead of first because it involved more radical departure in the beginning from the pure production rule system.

Instead of having this picture of production rules where you scan the data base to check whether there are any of the hypotheses satisfied and then put the conclusions in if they are you go backwards. You start with a conclusion that you would like to get to. You scan the data base. As you go through your production rules you scan the data base and see whether the conclusion in your production rules matches the things you're looking for in the data bases. The data base is now for the moment what you would like to get to instead of what you're starting with, so you just use the same system going backwards, and eventually you can make the same kinds of arguments.

That's a departure that's not quite as unstructured as the pure production rule system because you have to make some programming decisions about what conclusions you're going to try to draw.

The second way in which they differed from the PROSPECTOR people was that from the outset they decided they weren't going to try to do anything that was Bayesian, that could be justified in Bayesian terms.

They changed the words and instead of talking about probability they talked about certainty factors so they had certainty factors in the things they kept in their data base, and they made up their own calculus for the certainty factors.

Well, the interesting thing from the point of view of belief functions, of course, is these rules they made up turned out to be very close to the rules for belief functions, very close to special cases for demonstrations of the rules of combination for belief functions.

Eventually these folks got very interested in belief functions and looked at the cases where their rules differed from the belief function rules. In general, they seem to have drawn the conclusion that the belief function rules were better, and that they should recast their system in terms of belief functions.

I'm referring here to some recent work. Shortliffe was the guy in whose thesis MYCIN was originally developed. Gordon is a student of his and they have recently written a couple of papers in which they pushed the belief function idea.

In a particular context, they say that when they look back at their original diagnostic problem (they were dealing with medical diagnoses), when they look at the particular diagnostic problems they were considering they've decided now that those problems really had more of a hierarchial structure than they used to think they did - hierarchial in the sense that diseases form a diagnostic tree. I think you get the idea without my talking very much. (see Slide 11)

You can break a general disease down into more specific diseases, and then into yet more specific.

For instance, jaundice might be broken into a kind of jaundice that comes from an intrinsic liver problem or comes from some problem outside the liver; intrinsic liver problems like hepatitis or cirrhosis, things outside the liver that could cause the liver to malfunction or gallstones or problems with the pancreas.

They had the idea that perhaps the kinds of evidence they have could be dealt with in terms of belief functions and those belief functions would represent evidence that directly supported or refuted certain subsets in this tree.

The idea is you might have some test that told you yes or no the jaundice does seem to be the result of an intrinsic liver problem. That would indicate evidence directly for this hypothesis which is equivalent to the set consisting of these two hypotheses. The belief function idea is appealing because with belief functions you can represent having certain support for something here, without being specifically for either of these.

The hierarchial tree business corresponds to the fact they don't think that you're likely to get anything specifically that would support the subset consisting of cirrhosis together with gallstones. They don't see how you would get that kind of medical evidence because those two don't go together naturally.

So in general, if you have a tree the elements at the bottom are your sample space or your frame of discernment. You want to have the final probability judgments on subsets of the terminal nodes but only some of those subsets correspond to intermediate nodes. They think that you would start with belief functions that were focused on these intermediate nodes and then you could combine them by Dempster's Rule, et cetera.

Well, I have two comments on this. One is that Gordon and Shortliffe are very concerned about the computational complexities that result from Dempster's rule when you do this. They make some suggestions for modifying Dempster's rule and they get something computationally more feasible.

I think we can avoid that. I think that if we do exploit the tree structure we can find ways to efficiently calculate the probabilities or the degrees of belief that we need.

It's true that if you try to calculate degrees of belief for all subsets down here, you get into something that is unfeasible. If you recognize that what you really want is just degrees of belief for the subsets generally in the tree which do have high degrees of belief and you want to identify those subjects with high degrees of belief and find out what those degrees of belief are, I think you can do things efficiently from a computational point of view and you can use Dempster's rule.

I think the second thing to say is that in this picture you are likely to really need models for dependent evidence.

Considering something that might give you a medical test result, a medical test that might give you some positive evidence that the liver is involved directly compared to some other test that might be negative specifically for pancreatic cancer, the uncertainties involved in those two tests may not be independent.

So it's likely, it seems to me, that if you have a large number of items of medical evidence you're going to need models for dependent evidence in this picture as well as just the model for independent evidence that the Dempster's Rule is corresponding to.

Also I would ask the same questions here as I did in the case of the PROSPECTOR work. After you go this far and you find out that you're really dealing with hierarchical hypotheses, you find out you want to combine belief functions, you find out that you're going to have to have some models for dependent evidence, what's happened to the modularity of your knowledge and the flexibility that you started with when you were dealing with the production system idea? I think you've gotten quite a ways away from that.

My general conclusion from this picture - it's a little harsh - but I think it's true that the effort to put probability into production systems failed.

The reason it failed, one way of putting it, is that chunks of probabilistic knowledge are bigger than individual production rules.

That's not as revealing, I think, as this way of putting the problem. The point I was making when I was talking about canonical examples and constructive probability involving comparison to canonical examples is that probability judgment always involves design, because it does involve an overall comparison of your problem to canonical examples.

Another conclusion, is that probability judgment in expert systems is much like probability judgment on other problems, that the same kinds of difficulties arise. I think you get the same kind of contrast between Bayesian and belief function arguments. I haven't had time to go into that in terms of examples.

In those little examples I was showing you, we saw an advantage of belief functions over Bayes in that with belief functions you could use less complete models. I think that the same kind of flexibility is present in the artificial intelligence problems and does give a reason for preferring belief functions in some problems.

At the same time, in those little examples we saw that the belief function arguments sometimes do require models for dependent evidence. That same kind of thing arises with the expert systems problem.

Another conclusion I think we have to draw is about the future of expert systems that use probability. I don't think we can expect it to involve distinctively artificial intelligence ideas.

Though the term "expert systems" began in artificial intelligence, once they defined an expert system as a system that can have expert capabilities, it became clear there are a lot of systems that began outside the artificial intelligence community that are fully competitive in that respect. Artificial intelligence clearly has a problem now in redefining itself and deciding what part of expert systems is going to remain a part of artificial intelligence.

Having made these negative comments about artificial intelligence, I'd like to end with the comment that I think we should be very interested in what could be done with probability in more genuine artificial intelligence problems. As I said, apparently the reason artificial intelligence has gotten interested in probability was because they got away from the picture of nonnumerical inputs because they were talking about systems with numerical inputs from humans. But leaving aside expert systems, what about the argument that there shouldn't be any probability in artificial intelligence if you're dealing with nonnumerical input?

I think the field of artificial intelligence has outgrown that argument. I think artificial intelligence is no longer defined in terms of its contrast with number crunching.

There is a lot of sophistication about the levels of explanation idea, that you could use probability ideas even though you weren't starting with numerical inputs. You could imagine a system, like people, which generates probability estimates itself and uses those and combines them in various ways. Though I don't have time to talk about it and I wouldn't have very much to say if I did, I think an important area for people in probability to think about is the genuine artificial intelligence problem: how do you use probability in a genuine intelligence system that uses nonnumerical inputs, and where do the numbers come from. I think the answer is in what the psychologists have given us, some of the answers in the kinds of heuristics that humans use and presumably machines would have to use, too.

Where do the designs come from? Here I think we have to have bigger modules than production rules but there is still a lot to work on and think about.

Thank you all.

Example 1

Is Fred going to speak truthfully or carelessly?

S = {truthful, careless}

Are the streets icy?

T = {yes, no}

Experience gives me a probability measure on S:

P(truthful) = .8

P(careless) = .2

(Slide 1)

Bel(B) = P{s | if s C T, then T ∈ B}

$$
\begin{tikzcd}[column sep=small]
S = \{ & \text{truthful} \arrow[d, ".8"] & \text{careless} \} \arrow[dl] \arrow[d, ".2"] \\
T = \{ & \text{yes,} & \text{no} \}
\end{tikzcd}
$$

Bel(yes) = .8

Bel(no) = 0

(Slide 3)

Is Fred going to speak truthfully or carelessly?

S = {truthful, careless}

Are the streets icy?

T = {yes, no}

Fred says, "The streets are icy."

This creates a compatibility relation between S & T.

- truthful is compatible with yes but not with no

- careless is compatible with yes and with no

(Slide 2)

Bayesian Analysis

S = {truthful, careless but true, careless but false}

$$.8 \qquad \underbrace{.2q \qquad\qquad .2(1-q)}_{.2}$$

T = {yes, no}

$$P \qquad (1-P)$$

Form product measure eg $S \times T$, the condition.

(truthful, yes) (careless but true, yes) (careless but false, no)

$$.8 \qquad\qquad .2qp \qquad\qquad .2(1-q)(1-p)$$

$$P(\text{yes}) = \frac{.8p + .2qp}{.8p + .2qp + .2(1-q)(1-p)}$$

(Slide 4)

EXAMPLE 2

COMBINING EVIDENCE:

(A) FRED'S TESTIMONY YES .8

(B) THERMOMETER NO .99

s_1 = {TRUTHFUL, CARELESS}

T = {YES, NO}

s_2 = {WORKING, NOT}

$s = s_1 \times s_2$

= {(TRUTHFUL, WORKING), (TRUTHFUL, NOT),
.8 x .99 = .792 .8 x .01 = .008

(CARELESS, WORKING), (CARELESS, NOT)}
.2 x .99 = .198 .2 x .01 = .002

(SLIDE 5)

S	ELEMENTS OF T COMPATIBLE WITH S	PROBABILITY INITIAL	POSTERIOR
(TRUTHFUL, WORKING)	-	.792	0
(TRUTHFUL, NOT)	YES	.008	.04
(CARELESS, WORKING)	NO	.198	.95
(CARELESS, NOT)	YES, NO	.002	.01

BEL(YES) = .04

BEL(NO) = .95

DEMPSTER'S RULE OF COMBINATION

(SLIDE 6)

EXAMPLE 3:

DEPENDENT EVIDENCE

3 JUDGMENTS:

- FRED 80% RELIABLE
- THERMOMETER 99% RELIABLE
- FRED HAS 90% CHANCE OF BEING UNRELIABLE IF THE THERMOMETER IS NOT WORKING

S	ELEMENTS OF T COMPATIBLE WITH S	PROBABILITY INITIAL	POSTERIOR
(TRUTHFUL, WORKING)	-	.799	0
(TRUTHFUL, NOT)	YES	.001	.005
(CARELESS, WORKING)	NO	.191	.950
(CARELESS, NOT)	YES, NO	.009	.045

(SLIDE 7)

INDEPENDENT CASE (DEMPSTER'S RULE)

BEL(NO) = .95

BEL(YES) = .04

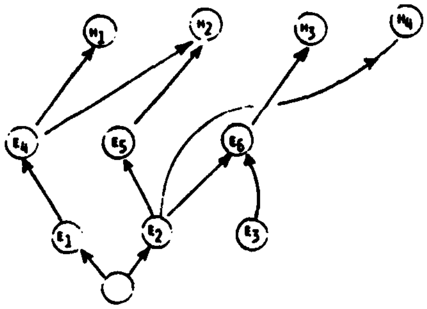
DEPENDENT CASE

BEL(NO) = .95

BEL(YES) = .005

(SLIDE 8)

PROSPECTOR



PRODUCTION RULES CORRESPOND TO LINKS.

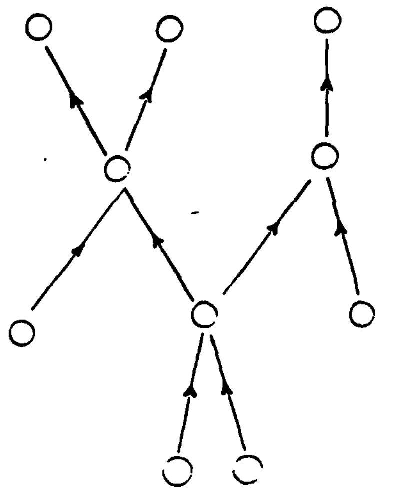
CONDITIONAL PROBABILITIES?

ALSO NEED PRIOR PROBABILITIES.

OTHER PROBLEMS: 1) CONDITIONAL PROBS MAY BE INSUFFICIENT
2) MAY BE INCONSISTENT
3) CYCLES?

(SLIDE 9)

JUDEA PEARL (UCLA)

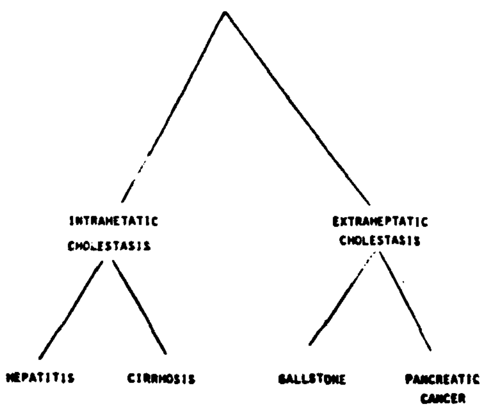


(CHOW TREE)

(SLIDE 10)

GORDON AND SHORTLIFFE

1984, 1985



"HIERARCHIAL HYPOTHESES"

(SLIDE 11)

DISCUSSION ON PRESENTATION OF GLENN SHAFER

DR. DeGROOT: There are two invited discussants for the papers at this time, Professor Art Dempster from Harvard University and Stephen Watson from Cambridge.

I'd like to call on Art Dempster for his comments.

DR. DEMPSTER: I may not be the best person to open up the discussion here, since I approach it from a very different point of view from most people who probably have many questions about what Glenn has been talking about. I hope there's lots of time for those questions to come up and clarifications to be made.

DR. DeGROOT: You don't have to use your full ten minutes.

DR. DEMPSTER: I probably will.

(Laughter.)

I think my role may be a little more to reinforce some of the things that Glenn has been saying and to complement a few of them by throwing out a few different kinds of ideas.

One of the things that we should be trying to do at this conference is bridge various language gaps. Different fields, engineering and AI and statistics, do tend to speak different languages even when talking about the same things so perhaps we can learn each other's language to some extent.

One of the things that to me is most appealing about probability and belief functions is that behind it are some very nice, straightforward, extremely simple mathematics which provides one with a calculus, so one can operate within a mathematical framework which is perfectly definite.

That's true of the Bayesian system and true of the generalization or the weakening to belief functions. For me, my own motivation in all this is much more to be moving toward practical things, moving toward specific models that people can use.

I've always felt that Glenn's approach is a marvelous idea, but we need to get in there and develop models.

I think one of the nice things about the belief function approach to expert systems is that it's a very rich field of application and one can do complicated things or simple things. The applied side of it, as I understand it, is really just developing.

So I am interested and have gotten more interested in pushing research in this field. I have the advantage of a very good student who is now working on it, Augustine Kong, who is here in the room someplace. I'm sure Augustine would like to tell us about some of his things if there is time or it's appropriate.

One can take some of these tree structures that Glenn has been talking about, some that don't have as many restrictions on them as the Chow trees and so on, and can develop belief function arguments and models and try to work out some of the difficult computational problems associated with it. That's the sort of approach that Augustine and I are trying to develop.

There are major technical problems there, specifying models and developing algorithms that can work in any kind of realistic computer time. Only then will we be able to test these models, test the systems, get some feeling for whether they can stand up under criticism and so on. For me it's not so much a matter of getting the axioms right and the calculus right and so forth. That's sort of all there. We need to use it and get some experience with it.

One thing that Glenn mentioned at the beginning was he was wondering why wasn't probability in AI, say from the beginning, or much more. That prompted me to think that, well, really probability isn't in anything much. The educated people, even the most technically educated people, generally don't think in terms of probability.

I spent a week last summer reading in economic theory and trying to understand what economic theory had to do with economics. I don't remember much about what I learned but I came away with one very strong impression. That although basic to the theory was the idea that people are out there with their own expectations and they're maximizing them, you would never find the word "probability" in the index. No thought at all where probabilities came from which underly these things.

In one field after another it's like that. In statistics, since mathematical statistics is so-called objective probabilities (which from one point of view are just kind of a way of avoiding the whole idea of probability), so AI isn't different in that regard. I think the practical end of AI is, as Glenn said, forcing some real concern about probability.

I think the statistics profession is to blame to some extent. It's *not* just that people aren't familiar with statistics, but there's a lot of confusion among statisticians which proceeds to their downplaying probabilities.

One of Glenn's themes is that AI is not necessarily the home for things like expert systems, or at least completely.

One thing that I've noticed starting to read into some aspects of AI, especially in the Bayesian area, is that there are a lot of parallels between the way statisticians look at a problem in some overall way in the way these people are doing it.

They're talking about levels of analysis. They start out doing things that we would think of in statistics as being exploratory data analysis at a low level; what is being perceived and recorded and so on. Then they have higher levels where there are references back to knowledge bases and things of that sort, which are much more in the spirit of doing Bayesian inference with models that have been established with exploratory or other more intuitive ways of doing things.

The kind of methodology that's evolving there has its parallels in the field of statistics. Another field that is involved is the field of design because there are always questions in expert systems as to what knowledge you are going to go out and try to bring into the analysis.

There is kind of unfortunate double use of the word "design" here. When Glenn uses "design," that's a term that I think came up in his work with Tversky, which is designs for heuristics in making probability judgments.

The other kind of design, designing what to pull out of the available knowledge with limited resources for collecting data, that's something that there's a lot of work done in statistics which should in principle relate to these expert system applications.

I'm not as big on heuristics as Glenn is. I have kind of an instinct that our knowledge has to come from empirical bases to a large extent if we want to communicate with each other. I do look to this kind of integrated picture of statistics as data analysis, and modeling, and cycling back and forth, and doing Bayesian inference in all of these things.

I think that that is the main source of the probabilities that we're going to want to use in formal statistics.

Let me stop there.

(Applause.)

DR. DeGROOT: Stephen Watson.

DR. WATSON: I don't want to take up too much of your time because I think most of you will want to be talking yourselves about these issues. I think the role of a discussant is maybe to be provocative and say things which people would disagree with more than anything else.

Ever since I've come into touch with belief function theory I've found it fascinating and a very interesting new development. It led me to come to lots of different conclusions. Several of them were reinforced this morning by Glenn's talk, which I thought was extremely interesting, and full of things to talk about.

There are just two points, however, I would like to share with you this morning concerning belief functions and their use in AI systems. These are the philosophical support for the theory and the notion of independence in the theory, which seem to me to be crucial to our understanding of this theory.

I take one of the roles of this conference to be a discussion of what different kind of calculus one ought to use in expert systems. It's clear there are competing claims for this. There is the Bayesian claim which Glenn talked about and the problems with that. Then there may be an alternative, or maybe it's just a different gloss on the same thing, the belief function theory.

One of the points which Glenn made in the paper that he prepared for this conference, concerns the philosophical support for the whole idea of belief functions, and related to that the philosophical support for probability theory.

Now there are some people, some in this very room, who will say that the only logically supportable way of handling uncertainty, wherever it appears, is to use the concept of Bayesian probabilities. That what we must do is to attempt to see how we can get over the problems of complexity which arise in trying to apply them in practical AI systems. That any other theory, fuzzy set theory, belief function theory or whatever, is philosophically unacceptable.

Now I think that's an unacceptable view. It seems to me that the philosophy of subjective probability is at best an article of faith.

There are lots of reasons to suggest that we as individuals do not handle uncertainty in our own minds according to the rules of the Bayesian calculus. What I find so exciting about belief function theory is that, as I understand it, here is another framework for thinking about uncertainty, looking at a different aspect of the way we naturally present uncertainty.

However, it leads me on to suggest that what we academics need to do is develop a new philosophy which relates to the belief function idea, rather than starting out with the philosophy of subjective probability as being the given as our basis for understanding these things. That's the first, I hope, contentious point.

The second point concerns independence. I suspect that at a conference like this there will be some people who know Shafer's theory intimately, and others who have heard and are interested in it but haven't really studied it. I fall somewhere between those two extremes. One of the things which grabbed my attention in studying Shafer's theory is, as I understand it, the concept of independence is not particularly

well advanced.

Glenn may disagree with me here and he knows better than anybody else how well the concept of independence has been advanced.

Simple applications of the Dempster rule for combining evidence, as Glenn said, only apply to independent pieces of evidence.

The question is when are two pieces of evidence independent and when are they dependent. If they are dependent, what do you do?

I think there is as yet no satisfactory theory for dealing with that part of the subject.

DR. DeGROOT: Let me open up the proceedings now to discussion from the floor.

DR. SINGPURWALLA: I'll raise a question both to Glenn Shafer and to Steve Watson.

Glenn mentioned the concept of independence. To me, independence can only be understood within the notion of subjective probability. When you refer to independence, what is it that you have in mind?

DR. DeGROOT: You want to take a second to answer that? I think it would be helpful.

DR. SHAFER: I don't know how much I can say that's useful but the general philosophical point of view that I was trying to advance a while ago is that what we're doing when we're making probability judgment is that we're comparing an actual problem to a picture of games of chance.

What you're saying is we do understand the picture of games of chance. You say independence is a probabilistic notion. Another way of saying that is what we do when we're talking about dice or perhaps about protons and photons; we do know what independence means.

DR. SINGPURWALLA: I can only understand independence using the calculus of probability. Is that what you're referring to when you say independence?

DR. SHAFER: Yes, and no. I am referring to independence as we understand it in the calculus of probability. If you're saying you're taking a practical problem where the calculus of probability is not there yet but you're making probability judgments, you're constructing a probability argument. You're comparing the practical problem to the picture of chance. Then you seem to have to make some kinds of intuitive judgments that that picture of independence fits.

DR. DeGROOT: It sounds like a topic that may be a recurring event in this conference.

DR. SINGPURWALLA: Can I make another comment, please?

DR. DeGROOT: You're paying. Go ahead.

(Laughter)

DR. SINGPURWALLA: The other thing you mentioned pertains to those hierarchial trees. I just want to draw your attention to the fact that in reliability we draw what are called fault trees and trace, because the events of interest are failures, the causes of failure. The calculus of probability has been used there quite satisfactorily. The main difficulties there happen to be computational. The tree gets very big and the question is how much time does it take to compute the top event. Of course the question of dependence and independence (as I understand it) arises there too; for that we use probabilistic models for dependence, and these are not readily available. There are of course few models for dependence.

DR. SHAFER: My only comment is yes, it seems to me those are problems. There must also be a problem about whether you know enough to fill in the details in all the conditional probabilities and probabilities in the tree.

DR. SINGPURWALLA: That is the difficult part. We have to assign conditional ones.

The idea is that you start with a very, very low level in the hierarchy where you can assign probabilities based on whatever reasons that you may have, experimentation etc. Then you build up the conditional probabilities and you build up the higher level probabilities by using, again, standard calculus of probability.

The Rasmussen report is an example of where this kind of thinking was used. I'm wondering if the artificial intelligence community and the expert system community has seen those kind of things, or have those kind of issues entered into that particular scheme of things?

DR. SHAFER: I think, yes, certainly that is a widely applicable idea, as applicable to diagnosing what's wrong with a car as it is to diagnosing what's wrong with the human body.

I think the use of those diagnostic trees is fairly late in this artificial intelligence story I'm talking about. That hasn't been a central theme; its only in some of this recent work of Gordon and Shortliffe where that came out.

DR. DeGROOT: Let me just interject one technical question that I have.

It seemed to me in your definition here using this compatibility relationship that the belief function would include those S's that were not compatible with any T's under your definition.

Does that make sense?

DR. SHAFER: Right. Part of the definition of compatible relations should be that for every S there is a T that's compatible with it.

When you do the rule of combination that gets violated and that's why you're going back and eliminating some of those S's by conditioning on the original space S.

DR. WISE: I'm Ben Wise of Carnegie-Mellon University.

When Steve Watson mentioned philosophical underpinnings of Bayes versus belief functions, one of those underpinnings is decision theory and how if you actually use a Bayesian decision rule you do minimize your expected loss.

I was wondering do you have any analog to decision theory based on belief functions and an argument coming out of that, that decisions based on belief functions will actually be good.

DR. WATSON: Could I just interject there and say that when I talked about philosophical underpinnings I was thinking of philosophical theories of probability particularly and not a decision theory concept based on expected loss.

Expected loss is either a notion which is ad hoc or it's derived from utility theory and probability theory which is axiomatized in lots of different ways, and I was thinking of an axiomatization for belief function theory or any other theory which might be different from that of probability theory.

DR. SHAFER: I do think that's a very interesting area. I could refer you to an unpublished paper that I've written on the subject if you're interested, but mainly it's concerned with the critique of the Bayesian underpinning decision, of the Bayesian utility justifications, rather than on any very great progress in a positive direction.

DR. WISE: If I may elaborate, do you have an idea of even how you would evaluate a decision rule based on belief functions if you can't use expected loss?

DR. SHAFER: Well, if you look at the mathematics of belief functions, the probabilities are going on in the space in the background, and the points in that space in the background are translated as subsets in the space in the foreground.

Now if you want to mesh that with a utility idea, I think the natural question to ask is why should the utility be attached to points in the space in the foreground?

Perhaps the utility itself should be attached to subsets. I think that you'll get a more interesting mesh if you take that approach.

It does seem to me that there is very strong argument against Art. Thats the line that kind of interests me. I don't know how far we can go with that.

DR. DeGROOT: Again, I think this raises an issue that will come up many times today and tomorrow, namely what is the operational meaning of belief?

My own view is that the purpose of an expert system, or indeed any information processing system, is ultimately to make decisions and I know how to use probabilities in decision-making. I don't know how to use beliefs in decision-making.

Again, I think that will be undoubtedly expressed again and again through the next few days and we'll hear, I hope, many different and interesting answers to that.

Are there other comments? David Spiegelhalter.

DR. SPIEGELHALTER: Just a note, really, just to point out what was said about the demands of probability may lose the modularity production rule. I think that that has also been realized that that's not suitable by people working in artificial intelligence on matters not concerning probability but on matters of explanation and control. The idea of having loose production rules which you can just plug in and take out has largely disappeared and much more structured knowledge bases now are becoming the norm.

DR. SHAFER: We see that's true of McDermott and people that are pushing like the XCON or DART. This may not be the best thinking that's going on in artificial intelligence, but it's still a very strong strand.

DR. SPIEGELHALTER: I was thinking much more in terms of the structure in CADUCEUS and structures in MYCIN and the control of the meta language, meta control that the Stanford group now has gone over to largely and the idea of a totally modular unstructured system.

DR. DeGROOT: Professor Zadeh, do you have anything?

DR. ZADEH: I'd like to comment on a question that was raised by Mr. Wise and that is that personally I prefer to use the terms upper and lower probabilities to belief and plausibility.

I think that when you associate the term belief with lower probability, you tend to read more into it than really is justified, so that if you look at it in terms of upper and lower probability, if you look in terms of probability bounds, that's basically the point if you take in Dempster's original paper.

Then the answer to the question that was posed is that correspondingly then you will have upper and lower expected values. That's what you will have. In other words, whatever values you compute based on incomplete information will be in terms of bounds and it will be then up to the decision analyst to decide what to do, if all you know is that the expected value lies between alpha and beta.

DR. DeGROOT: My thought about that comment is that the argument against probability is that it's so difficult to specify precise probabilities that one really can't use them in a practical way in large-scale and important problems. The suggestion therefore that they should be replaced by upper and lower bounds seems to be saying it's too hard to specify a single number, therefore we'll specify very precisely two numbers -- an upper and lower bound. I find that very difficult to accept.

DR. YAGER: I just want to make one comment about that idea of modularity you were talking about - losing modularity.

Perhaps what may happen is that you'll have -- instead of it being totally modular as it is now, you may have some chunks of information that are sort of clumped together by some sort of probabilistic information so instead of the whole thing being independent you'll have sort of groups of information being independent.

DR. SHAFER: Spiegelhalter may have more to respond to that than I do.

DR. SPIEGELHALTER: Those are highly structured groups of maybe up to ten nodes with full probability distribution defined on these, and yet each of these is essentially independent.

DR. DeGROOT: Well, it is time for coffee, I think. I would like to thank everyone. It seems to me that one conclusion I've gotten out of this is that adherents of the Bayesian approach are called Bayesians and to them the rest of the world is non-Bayesian, and I gather that adherents of belief functions are believers and the rest of the world is nonbelievers. (Laughter).

(Recess.)

FUZZY SETS AND POSSIBILITY THEORY

Lotfi A. Zadeh
Computer Science Division
University of California, Berkeley

Presentation was based on material published previously:

L. A. Zadeh, "The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems," Fuzzy Sets and Systems II, (1983) 199-227.

L. A. Zadeh, "A Simple View of the Dempster-Shafer Theory of Evidence," Berkeley Cognitive Science Report Series, University of California, Berkeley (1984).

L. A. Zadeh, "Fuzzy Sets and Information Granularity," Advances in Fuzzy Set Theory and Applications, M. M. Gupta, R. K. Ragade, R. R. Yager (editors), North Holland Publishing Co., (1979).

TRANSCRIPT OF ORAL PRESENTATION BY LOTFI ZADEH: FUZZY SETS AND POSSIBILITY THEORY

DR. ZADEH: To place what I have to say in the proper perspective, let me say something about my perception of how the theme of this workshop fits into some of the issues in the case of expert systems.

If you look at the dilution of scientific progress in various fields, you observe the following. You always start with a deterministic theory. Gradually the realization develops that knowledge in that field is not really deterministic. The next stage is to go to a probabilistic description. Eventually it turns out that your knowledge of probability is incomplete. So then you add incompleteness to it. And this is really the situation we have in the case of expert systems. The knowledge that is in the knowledge base of an expert system typically is incomplete.

The question then is how do you come to grips with the problem of incompleteness. The Bayesian approach, as I understand it, is to say that there is no such thing as incompleteness.

Professor Lindley disagrees with me. Perhaps he might correct me on this point. You assume that you can always make up for incompleteness through the use of subjective probabilities. That is one point of view.

The point of view that is taken in Dempster-Shafer theory is that you do have incompleteness in your knowledge of probabilities. This is what Glenn alluded to in his statement that you deal essentially with an incomplete model. I think Dempster used the word "weakening."

When you see this sort of a thing the incompleteness in your knowledge propagates down to your conclusion, as a result of which the probability has become interval-valued in some sense. So you have to speak about the lower and upper probabilities, or belief and plausibility.

Another approach is based on the use of the maximum entropy principle. Here I do have incomplete information, but I am going to make up for it by making certain assumptions, which, in some unsophisticated way, are assumptions about independence.

Not knowing what the joint probability is you assume that it is essentially the product of the margins, or roughly like that. Of course I am oversimplifying things. You make up for it in that fashion.

Personally I feel that the spirit of the Dempster-Shafer theory is proper. That is, you should not really make up for incompleteness by making all sorts of assumptions, regarding either independence or assumptions about probabilities that you really don't know.

It is at this point then that Possibility Theory enters into the picture. Possibility Theory is essentially a theory of incompleteness. Let me try to explain what is involved by starting with a very simple thing.

Suppose you have a variable X. You say X is equal to A. You assign a value to X. A weaker statement might be that X belongs to some set A. Now, when you say that X belongs to A we are making a possibilistic statement.

We are saying that the possible values of X lie in this set. But this sort of possibility is the 1-0 possibility. It is the all or nothing possibility.

The possibility that is used in possibility theory is a matter of degree. Generally it would be associated with statements of the form X is A, rather than X belongs to A. For example, X is small; X is large; Mary is tall; and so forth. Statements of that kind are possibilistic statements, but in this case possibility is a matter of degree.

We can talk about physical possibility; for example, the possibility that you might lift 50 pounds, 100 pounds, 150 pounds and so forth. As you start with zero pounds and go on to 500 pounds, there is a gradual transition from being able to do it quite easily to not being able to do it at all. Notice that when you talk about possibility in this sense, there is nothing probabilistic about it. It is simply a matter of ease of attainment. But you are talking at this point in terms of physical possibilities.

Consider possibilities that are associated with statements or propositions of the form "X is small." The interpretation that you put on that is that "small" is something that can be stretched. In talking about the degree of stretch, if I say Mary is young, and Mary in fact is 35, then you have to stretch the concept of young to a certain degree to accommodate the value 35. We have stretch but it is not physical stretch. It is a conceptual stretch. This is really what happens. As far as I am concerned then, the Dempster-Shafer theory is one in which you have some knowledge about probabilities and you have some knowledge about possibilities. It is a mixed theory. It is certainly a step in the right direction because it is a generalization of classical probability theory.

Fuzzy logic or fuzzy sets come into the picture in the following way.

In using fuzzy logic, you do not have this separation between logic and probability as we have in the case of classical logic. They are under the same roof. Here is a fuzzy logic.

First, you have the representational components. The representational component has to do with taking something that is expressed in a natural language and translating it into a more precise language.

It is the sort of thing that you do when you use predicate calculus; you take something and express it in the form of predicate calculus.

Then you also have an inferential component. The inferential component comes into the picture once you have represented knowledge, expressed in a natural language, in that more mathematical, more precise form.

How can you infer certain propositions from other propositions? Generally that requires the solution of a nonlinear program. What happens is that in classical logic we use tools such as resolution, things of this kind--modus ponens.

From the point of view of fuzzy logic, these classical rules are merely degenerate forms of nonlinear programming. If you take this point of view, you begin to see more clearly why things work the way they do.

Fuzzy logic subsumes probability theory through the use of numerical, or more generally, fuzzy quantifiers. In the case of fuzzy logic, for the most part you deal not with probabilities but with quantifiers, like several, many, most, few, and so forth which in some sense brings probability theory to the pre-Kolmogorov era.

In other words, this is the way probability theory was looked at many years ago, except that you incorporate it into this quantifier structure. Furthermore, you allow these quantifiers to be fuzzy.

The concept of a quantifier is closer to human intuition than the concept of probability. You really don't have to use probability. You can formulate, for example, Dempster-Shafer theories and other theories without ever mentioning probability theory. If I have a chance, I will show this later.

In the case of expert systems, much knowledge is the knowledge of "usual values." What one can do--and this is what I am trying to do at this point--is develop a theory perhaps called the theory of usuality.

The concept of usuality, the concept of usual value, differs from the concept of expected value. I think it is really more relevant to decision-making. Let me give a simple example of what I have in mind.

An expected value is simply an average value. If I tell you that the average value in some location is 70 degrees, it doesn't mean that much; it could be extremely hot in the summer and extremely cold in the winter.

But if I say that the usual value is 70 degrees it means much more because the usual value is really representative of what the value is, whereas an expected value is not.

So the question then is how does the concept of usuality tie in with expert systems, and how can one develop what might be called a calculus of usuality.

The concept of usuality is cause-related to another concept, that of disposition. A disposition is a proposition which is, for the most part (but not necessarily always) true.

The knowledge and the common sense knowledge that we have consists mostly of these dispositions. Furthermore, the concept of a disposition manifests itself in many other ways. Here are some examples.

As a proposition: slimness is attractive.

As a valuation: it takes about five minutes to reach the station.

As a command: avoid overexertion.

As a ranking: Swedes are taller than Greeks.

As similarity: Spaniards resemble Italians.

As causality: overeating causes obesity.

As typicality: a typical Swede is tall and blond.

As usuality: usually pork is much cheaper than veal.

What I am trying to say is that in our preoccupation with probabilities we have tended to lose sight of the fact that much of the information on which decisions are based does not really fit classical model probability theory.

Technically, disposition is a proposition which is preponderant but not necessarily always true. A disposition is a proposition with implicit fuzzy qualifiers.

For example: birds can fly; most birds can fly. Young men like young women; most young men like mostly young women. So what happens is that when we assert something that is a disposition, there are implicit fuzzy quantifiers.

You can interpret this quantifier in terms of probabilities, but I prefer not to do so and stay on the level of these quantifiers. What happens then is that if you take these dispositions and apply modus ponens or modus tollens, then you get certain decision principles.

For example, if A implies B, then to achieve B, do A. Slimness is attractive. To be attractive, be slim.

Notice that these things don't guarantee that if you achieve slimness you will also achieve attractiveness. This is simply a tendency, a disposition. Or you may have a negative decision principle. If A implies B, then to avoid B, avoid A. Overeating causes obesity; to avoid obesity, avoid overeating. What you have to do is make knowledge of that kind more precise.

Let's start with something like "slimness is attractive." That can be interpreted in a number of ways. For example, most of the slim are attractive. Most of the attractive are slim. There are many more attractive individuals among the slim than in the general population, and so forth. There are at least nine different ways in which you can interpret this. Any one of these nine or more different ways are admissible. In other words, if you want to say that I interpret it in the sense of No. 5, that is perfectly all right. The next question is how would this statement be interpreted or understood in usual conversation. Will it be sense No. 1, or sense No. 2, or whatever? There is the issue.

These fuzzy quantifiers that enter the picture are generally second-order predicates which characterize the absolute or relative count of elements in one or more fuzzy sets. In the case of fuzzy sets, you have a class that doesn't have sharply defined boundaries; therefore, it does not make sense to ask how many elements are in that set.

Nevertheless, you can take the classical definition of cardinality and extend it to fuzzy sets. The simplest extension and the one that is presented on the next transparency involves simply adding up the grades of membership of the elements in the set.

This is called the sigma count. If you have a fuzzy set--that is represented by these dotted lines--and if you have a point U and it has the grades of membership U in the set, you simply add up the grades of membership.

This is a little bit like full time equivalent. That is the way university administrators add the number of faculty and students and so forth. You can form the relative sigma count, which is the sigma count of the intersection divided by the sigma count of A.

If you say that QA's are B's, in effect you are saying that relative sigma count of B in A is Q. Notice that I am saying is Q not equals Q because it is understood that Q is a fuzzy number. That is why I am not equating the sigma count of B in A to Q. With this definition of a quantifier you can manipulate the fuzzy numbers. In other words when you say "most," you are characterizing in a fuzzy sort of way the proportion of elements of one kind in elements of another kind.

"Most," for example, then would be a fuzzy number like this. This fuzzy number is a possibility distribution. That is, if you say that something is "most," then this would present the possibility that that proportion has a specific numerical value.

For example, if .8 is somewhere here, you read this value here, and that will give you the possibility that it has that value. Once you have defined these as fuzzy numbers, you can manipulate them using fuzzy arithmetic. For example, you can square them. You can add them. You can divide them. You can do various things with these fuzzy numbers. You can also represent them as ultra-fuzzy sets so that in this particular case the possibility distribution in itself is also fuzzy.

It is at this point that the concept of usuality comes into the picture. Usually it is interpreted as a quantifier, in the following sense.

You say that "usually" X is F. What is implied is first of all, there is a conditioning variable Z. You say that if Z is not some exceptional cases or you make a positive assertion: it belongs to a certain set, which is the set of normal values, then, most X's are F's. That is the conditioned version. The unconditioned version is simply usually X is F, which is most X's are F's. Here is an example: it takes a little over an hour to drive from Berkeley to Stanford.

Now notice that as it stands, the word "usually" does not appear in there. It is implicit. So D stands for disposition and R here stands for restoration. You are restoring disposition if you make the quantifier explicit.

So you say what it really means--here it is unconditioned--is that usually that duration is little over one hour. "Little over one hour" is a possibility distribution. You are giving the possible values of that variable but this is not a probability distribution. The probabilities indirectly come in the use of the word "usually."

Now you can condition that. You can say that if departure is not rush hour, then usually duration is a little over one hour. And then you can define more precisely what you mean by using what is called test-score semantics.

Test-score semantics tells you the following. It says I will be able to tell you the degree of agreement of that proposition with what is in the data base if you tell me the entries in the data base.

The meaning itself is the procedure. It is not the values in the data base. That is why this data base is called an explanatory data base. It is something that you construct for purposes of explanation.

In effect you are saying that if you gave me a record in which you have trip one, trip two, trip three and so forth, and these things here--point 8, point 3--this is the degree to which the duration for that trip agrees with little over one hour, in this case you will agree to degree .8, .3, and this is the degree to which the time of departure agreed with the constraint "not rush hour."

If you give data like that, then by going through this computation which also involves definition of "most" (I will not go into the details) you will be able to compute the degree of agreement.

It is that computation that defines the meaning of the proposition which is expressed in natural language.

DR. GROSS; What does the .8 and .3 actually mean when you say "the agreement?"

DR. ZADEH; This is a question which is something that most people raise. In other words, how do you get these numbers? The assumption here is that these numbers are your perception of the agreement. It is the sort of a thing that humans are good at but we do not understand too well how we do it. Now frequently you are asked to fill out a questionnaire, or write a letter of recommendation, or whatever. In filling out this questionnaire they have a certain scale--on a scale from zero to ten indicating degree to which this student is outstanding or whatever, indicating something.

People don't have too much difficulty in doing that without really understanding how we do it. Olympic judges do that. This is a basic issue. I am not trying to minimize its importance but for the moment I want to put it aside.

Let's assume that in one way or another if somebody asks you the question "it took an hour and a half to get from Berkeley to Stanford; to what degree does this 1.5 hour agree with your perception of about one hour'?"

So you will say .2, .3. That is how these numbers are obtained.

What happens is this: The contention here is that what we call the knowledge base consists really of the knowledge of usual values. And notice one thing, that if we did not know what the usual values are, you wouldn't be able to do a thing.

Now the reason we can function is because you know that it takes about one minute to get from here to the elevator; it takes about five minutes to get from here to something. It takes about three dollars to have lunch in this cafeteria. You have this tremendous store of information about the usual values, not just of various parameters, but also usual values of relations. For example, it may be that something is much larger than something else, and so on.

When you say the usual value of X is F, what you mean by that is that usually X is F. When we have usual values of a pair (X, Y), that is a relation really. That means usually X,Y is R. For example, usually X is much larger than Y. Usually X is small. Usually most X's are small and so forth. Now the usual value of X, as I have indicated already, is not the expected value of X.

Now what matters in decision analysis is the usual rather than expected value of X.

DR. SHAFER: Does it very often matter what the usual value of the distance from the usual value is? Do you want very often to take into account how far you usually are from the usual value?

DR. ZADEH: Yes. I will come to that in a moment. Let's look at that a little more carefully. Perhaps that may respond at least in part to your question.

Suppose we have what is called disposition evaluation: usually X is F. What do we mean by that really? Now first of all, suppose that this F is about alpha. So here at this point you have a possibility distribution.

Now in special cases it would be an interval, for example between something and something. But this is now a possibility distribution. Is something small, large, young, something like that.

Now usually is also a possibility distribution. Here you have then a mixture of two possibility distributions, one defining usually and the other defining what the value of the variable is. When you say usually X is F, you have to interpret that. This is where the representational component of fuzzy logic comes in the picture.

It says that you should interpret it in the following fashion. If you knew the probability distribution on X, let's assume we have the simple probability density, then take the membership function that is associated with the distribution and calculate this integral which is the expected value basically in probabilistic terms of this characteristic function.

You substitute that into the definition of usually and you come up with a number. That number is the possibility of the probability density. It is a possibility of the probability.

In other words, it says that when you tell me that usually X is F you are giving me information about the possibility distribution of the probability distribution. That is what you are really telling me.

Now this is simply probability. Then there is the issue of informativeness. How informative is a statement like that? There are two things involved: one is the specificity of F. Specificity is a concept that Ron Yager has written on extensively. It is essentially how narrow that thing is, how restrictive it is.

Obviously you are not giving too much information if this thing is very broad. Nor are you giving any information if "usually" is very broad. So the informativeness of that piece of knowledge then is a function of the specificity of this and the specificity of that. (And for simplicity you can form the product if you want to define "informativeness").

So this is the way in which "usually" would be interpreted. Now how can you compute with this sort of a thing. Let's take a very simple example. Suppose that you know the usual value of X and you know the usual value of Y. What would be the usual value of X plus Y, the most elementary question you can raise. In the case of expectations, of course, we know it is A plus B. Now if you say for simplicity (let's assume that these are numbers), X is A and Y is B, then X plus Y is A plus B in fuzzy arithmetic.

Fuzzy arithmetic is a generalization of interval arithmetic. In the case of interval arithmetic, you are dealing with possibility distributions that are either zero or one.

When you say that the number is in this interval, you are saying that the possibility that it is there is one and the possibility that it is outside is zero. So this addition of fuzzy numbers is a generalization of the addition of interval-valued numbers. In effect you are saying then that X plus Y is this. What you can show is that usually X is A and usually Y is B implies that usually plus X plus Y is A plus B. Plus means it is a little bit narrower. How much narrower you can't tell without having more information about these things.

The statement becomes a little bit sharper. It is not just usually X plus Y, but it is usually plus X plus Y is A plus B. If you wanted to perform a more careful analysis of this, consider the following practical problem.

You have dinner at a restaurant. You have the cost of the appetizer; the cost of the entree; the cost of dessert. You have cake. So the total cost is expressed by this. You know the usual values--usually X is about $3, usually Y is about $10, and so on. What is the usual value for the dinner? If you take the same problem that I considered previously, Z is equal to X plus Y, instead of the kind of result in the previous slide (where you take the sum of A plus B and ask what quantifier can you associate with A plus B) you ask another kind of question.

You ask if I want to stick to usually, what is that value that can associate (not necessarily A plus B) with the sum. It is a different kind of question.

I will not go through this analysis but it is more complex. Again dealing with possibilities, it turns out that eventually you have to solve certain equations involving these possibilities and that in turn will require the solution of a nonlinear program. In general the nonlinear programs that result from formulations of this kind are continuous nonlinear programs. In other words, they are programs in which you don't have simply vector-valued variables, but rather the variables are probability densities and things of this kind. It becomes expensive to solve these problems. At this point you do not have quick and dirty ways of solving nonlinear program problems of that kind. This is a question that was raised by Glenn in his talk.

Is there a distinctive AI flavor, something other than straight probability analysis, in the case of expert systems? I think that this is in that spirit. It is not in the spirit of classical probability theory.

It is a mixture of logic together with fuzzy quantifier probability. This is a dispositional modus ponens and is really what is needed to be able to propagate this knowledge of probabilities upward on the trees or inference networks that Glenn mentioned.

Suppose you know that Q X is F. Now Q is some sort of a quantifier. It could be "usually" but it doesn't have to be. It could be "about 80 per cent of the time." It could be anything you wish.

Q is F. If X is F, then Q Y is G. Notice that as was mentioned already in the case of production systems, usually you have rules of the "if A, then B kind." If A, then B. And you also have facts about A.

Here that fact about A is dispositional, in the sense that there is a fuzzy value associated with it and also this Q. My contention is that this is really the kind of information that people have. This is the kind of information that geologists have. This is the kind of information that doctors have. Furthermore there are the rules.

Here the assumption is that you allow this quantifier, which could be interpreted as probability if you wish, on the right side of the rule. It turns out that the conclusion that you can reach here is that Q^2, (Y is G) is the square of that fuzzy number that is associated with Q. For example, usually X is F. If X is F, then usually Y is G. Therefore, Usually^2 (Y is G). This is the sort of conclusion.

Here is the picture. Usually^2 is less specific. As a result of the use of these chains of inference, the results become fuzzier and fuzzier and fuzzier.

Basically the conclusion that emerges is that, in general, these chains of inference cannot be long. In my scene they can be arbitrarily long but then the validity of the conclusion is very much in question.

Basically you can't use long chains of reasoning if your information is imprecise. Here is another rule. In fuzzy sets and possibility theory there is a rule that is called the compositional rule of inference.

The compositional rule of inference is shown here. It says X is F. X and Y are G. Therefore, Y is the composition of these two relations. F is a unary relation and G is a binary relation. Notice that there is nothing that is probabilistic about this sort of thing. Mary is tall. John is much taller than Mary. Therefore, John is the composition of tall and much taller.

How this is actually performed is shown over here. You have to work with the membership function. Then you take the supremum over this sort of a thing and then that is the result.

This may be viewed as the analog of the classical Bayesian rule; the probability distribution of Y is the convolution of the probability distribution of X and the probability distribution of Y given X.

In the case of fuzzy sets, instead of dealing with integrals, you usually deal with suprema and instead of dealing with products you deal with the min operator.

If you take the standard formula in probability theory--the probability of Y is equal to the integral or summation or probability of Y given X times the probability of X--this is its analog, the composition rule.

Then the question arises, what happens if you qualify these things probabilistically, or here in terms of usual values. That usually X is F and usually X and Y are G. You qualify these things. Again, this is like Glenn mentioned in his talk, that you take these rules and associate some probabilities for certain factors (or something like that) with these rules.

This is in that spirit. The question is can you say that Usually^2 Y is F composed with G. The answer at this point is--I am not sure. It looks reasonable but in order to justify this one has to go through a reasonably complicated analyses. I have some transparencies here but I will not show them and I have to solve some problems again in nonlinear programming.

It is possible that that result may be good; again I am not sure. Here is the nonlinear programming problem to which this reduces.

This here is greater than or equal to the supremum of this rather messy looking expression over here and we have to find the smallest membership function which satisfies that.

What happens is that in principle you can take problems like that and reduce them to the solution of these nonlinear programs. In practice, the stumbling block at this point would be how to solve those nonlinear programs in an approximate fashion, because what is going to happen is this. If you want to solve these nonlinear programs using standard software that is available for solution of such problems, you would be wasting the capability of a computer to solve these problems precisely because you are not really interested in the precise answers to these questions. You are interested in something that is about the same order of precision as the original data, which is not that high.

This suggests a view of decision analysis that is again different from the classical one, which might perhaps be called dispositional decision analysis.

In this dispositional decision analysis you assume that the values of various parameters are dispositional valuations. For example, alpha is a parameter. You say it is something which is a fuzzy value, but there is an additional quantifier (implicit quantifier) which is "usually."

"Usually" a cup of coffee costs about 50-cents. What happens is this: you can take some standard problems in decision analysis (this is one involving Markovian decision analysis) where you have a system with the number of states. You have transitions from states; each denoted condition probability for each transition; and you know also the cost associated with the transition. You want to find a policy, that is what input to apply when you are in a specified state, in such a way as to minimize the expected cost.

This problem has been considered extensively in the literature. By using dynamic programming it can be reduced to the solution of a functional equation. The dispositional aspect comes into the picture in the following fashion.

These transitions would be of the form which is shown here if the input is alpha K and you are in state QY; then the next state is "usually something." That is the way it would be specified. How can you find an optimnal strategy for a situation in which that is the kind of information that is available. It turns out that one can take this equation (which is the one that one finds in standard Markovian approches) and this can be solved by using fixed point methods. That is one way of solving it, regarding that as an equation of the form that X is equal to some function of X, and using fixed point iteration. This fixed point will now be a fuzzy set. It will not be a simple form. It will be a cloudy thing fuzzy set, but in the literature on fuzzy sets there are several papers which deal with the extension of various fixed point theorems to fuzzy sets, and those extensions may be relevant to the solution of this problem.

There is something that has been generating a lot of interest in recent years in AI, and this is the subject of non-monotonic reasoning. It relates to the issue of dispositionality. As far as I can see, what people call non-monotonic reasoning is simply probabilistic reasoning. It is not non-monotonic. In other words, I question the use of the term non-monotonic to describe that.

Here is a simple example of what people call non-monotonic reasoning: Slinky is a bird. What is in parentheses here is what i. implicit. Birds can fly. Therefore, Slinky can fly.

Again, here in parentheses you have some form of qualification, like it is very likely that Slinky can fly. Disregarding that, then you say Slinky can fly. Then you have additional information.

Slinky is an ostrich. Ostriches are birds. Ostriches cannot fly. Therefore, Slinky cannot fly. Now people say that this is non-monotonic reasoning because here you are of the conclusion that Slinky can fly and here you are of the conclusion that Slinky cannot fly.

In classical logic that cannot happen. In classical logic, as you add propositions to your premises, the truth value of a conclusion can never change. If it has been established to be true it will continue to be true.

The point I am trying to make is that if you look at this as probabilistic reasoning there is nothing that is nonmonotonic. It is simply a matter of revision of belief in the very classical sense. That is, here you have two pieces of information; that Slinky is a bird and Slinky is an ostrich.

But ostriches are subsets of birds. So you have the conditional probability of A, flying given that the bird is an ostrich, but that is the same as the conditional probability of A given B because B is a subset of C by the simple rules of conditioning.

Therefore, the probability that Slinky can fly is zero; Slinky cannot fly, because you know it is an ostrich. So the reason why people think it is non-monotonic is because those implicit probabilities are disregarded.

Once you have made them explicit, then the non-monotonicity disappears. One important issue that arises is how can you combine evidence using these quantifiers?

The basic idea is that you use fuzzy syllogisms. What is a fuzzy syllogism? It is something of the following form:

$$Q_1A\text{'s are }B\text{'s; }Q_2C\text{'s are }D\text{'s;? }Q_3E\text{'s are }F\text{'s.}$$

Now here are some examples. Most young women are slim. Many slim women are attractive. What fraction of young women are attractive? Now A, B, C, D, Q_1, Q_2 are all assumed to be fuzzy.

Now depending on how A and B and C and D are constrained, you get different syllogisms. For example, chaining involves a situation in which this B is the same as this C, this E is the same as A, and this F is the same as D.

Then you have consequent conjunction, antecedent conjunction, dissection product and so forth. You have a number of syllogisms depending on how these things are constrained.

The Dempster-Shafer rule of combination relates to just one of these syllogisms. In other words, you need not have just one rule of combination but many in order to be able to chain, to deal with disjunctions and various other things.

Here is an example of a syllogism in fuzzy logic. Q_1 A's are B's. Q_2(A's and B's) are C's. Therefore, (Q_1 times Q_2)A's are B's and C's, from which you can infer that at least (Q_1 times Q_2)A's are C's from which in turn you can infer that (Q_1 times Q_2)A's are C's, if Q_1 and Q are monotonic quantifiers or monotonic numbers.

This intersection product syllogism in probability theory would correspond to the basic formula in which you have expressed the probability of A given B and C or probability of B given C in terms of some of the constituent probabilities, so that if expressed in terms of quantifiers, it would be a very elementary rule in probability theory.

The difference is this. All of these things that appear there are allowed to be fuzzy, whereas in probability theory they are not allowed to be fuzzy. Events are not allowed to be fuzzy.

This intersection product syllogism then can be applied here to situations like this: Q_1A's are B's; Q_2(A's and B's) are C's. That is a syllogism. Now if you put (almost all)A's are B's and all B's are C's, then from that you can infer that all A's and B's are C's.

From these two now, you can infer that (almost all-times-all) A's are B's and C's and it is understood that this product is a product in fuzzy arithmetic rather than a product in ordinary arithmetic.

And (almost all-times-all) lacks unity, so "almost all-times all" is the same as (almost all)A's are B's and C's from which you can infer that at least (almost all)A's are B's and from which you can infer that (almost all)A's are C's.

From (almost all)A's are B's and all B's are C's you can infer almost all A's are C's. I should like to note the classical syllogism of Aristotle. All A's and B's, all B's are C's, therefore all A's are C's. That is a standard syllogism.

This is a variation on that where in the first premise you relax a little bit and make is "almost all A's are B's." Then it turns out that almost all A's are C's. However, if instead of introducing "almost all" in the first premise--the major premise--you introduce it in the minor premise, then this whole thing will collapse. In other words, it is brittle.

The difficulty then is that if at some point you replace (in terms of probabilities) probability one by probability one minus epsilon the whole chain of reasoning collapses completely. Whereas in other premises, it is okay. What complicates this and has not been considered in the theory of expert systems is that the place for introducing these probabilities is critical.

It is a little bit like what happens to round-off error in the course of performing numerical computation. In some places it is okay. In other places it can have disastrous results. This is the problem.

People assume that these computations are robust, but in fact they may not be robust. What happens is, for example: This is a consequent conjunction syllogism.

Q_1A's are B's. Q_2A's are C's. What is the fraction of A's and B's and C's? It turns out that there is a bound on Q. Now these are operations on fuzzy numbers and in particularly if Q_1 and Q_2 are most, then the bound on Q is two-times-most minus one on one side and most on the other side.

What happens is that this incompleteness of information translates into bounds and quantifiers. This is one of the serious weaknesses of these traditional approaches to dealing with uncertainties of expert systems. There is an implicit assumption of compositionality. That assumption is made in MYCIN. It is made in PROSPECTOR. It is made in all the systems that have been devised so far. That is, you assume that if you associate a certain factor with a certain rule, and another certain factor with another rule, and these two rules are combined, then the certainty factor associated in the combination would be a number.

This sort of a thing suggests that it will not be a number in general. Even if you start with two numbers to begin with, the result will be an interval valued number if you have no fuzziness in the picture.

If you do have fuzziness, then it will be a fuzzy interval valued number. So this compositionality then is lost. And because it is lost, the computations very quickly become uninformative. I think that the long chains of computations that are allowed in MYCIN are really not justified. There is another problem that I want to mention in connection with MYCIN. There is one serious flaw as far as I can see and that is the certainty factor is taken to be the difference between the measure of belief and the measure of disbelief. That means that you have certain supportive evidence which tends to lead you to the conclusion that the hypothesis is true; there is another body of evidence which tends to lead you to the conclusion that it is not true.

If you compute the two and subtract one from the other the net difference is taken to be the certainty factor for the conclusion. That sort of a thing can lead to the following highly counter-intuitive situation:

You have 100 witnesses testifying and 99 of them say that the defendant is guilty and one of them says that the defendant is not guilty. The one negative vote completely nullifies the 99 positive votes.

In other words, evidence is not cumulative in MYCIN. This is what happens, which is counter-intuitive. So the situation then is this as far as I can see. To summarize what I have said is that I believe that Dempster-Shafer theory is a very useful theory. It is a very interesting theory from the theoretical point of view and it is a very useful theory from the practical point of view.

It is certainly a step in the right direction. I think that by itself, however, it is not sufficient; that is, you have to have many other rules, of combination, for inference and so forth, in order to be able to deal with the problems that one encounters with inferential processes in expert systems.

I think probability theory by itself is likewise insufficient. That is, one has to have at one's disposal probability theory and possibility theory and use the two of them, generally in combination. In that way you arrive at answers whose precision is commensurate with the imprecision of the information knowledge base.

You do not have the kind of artificial precision that you get out of existing expert systems. MYCIN, PROSPECTOR and so forth give you numbers which are misleading because there is really no justification for that high degree of precision if you use any kind of established theory, be it probability theory or some other theory.

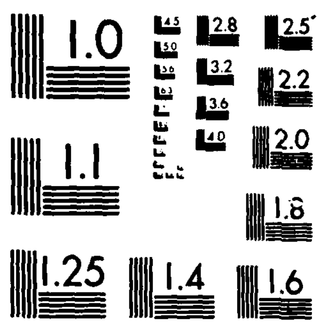
All of these theories will lead you to the conclusion that what you can assert about the certainty factor of the final answer is much less specific than the existing expert systems would give.

Thank you very much.

DR. DEGROOT: Thank you very much.

(Applause)

AD-A163 642 THE CALCULUS OF UNCERTAINTY IN ARTIFICIAL INTELLIGENCE 2/3
AND EXPERT SYSTEMS..(U) GEORGE WASHINGTON UNIV
WASHINGTON DC INST FOR RELIABILITY AND..
UNCLASSIFIED N D SINGPURWALLA ET AL. 15 JAN 86 F/G 9/2 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

MANIPULATION OF FUZZY QUANTIFIERS

80% of students are single

60% of single students are male

80% × 60% of students are single males

most students are single

a little more than a half of single students are male

(most)⊗(a little more than a half) of single students are male

(most)⊗(a little more than a half) = about a half

rounding to the nearest comprehensible linguistic term.

AΣ Count

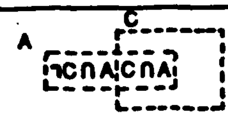
$$\Sigma Count(C/A\cap B) = \underbrace{\Sigma Count(C/A)}_{\text{is } Q_1}\ \underbrace{\Sigma Count(C/B)}_{\text{is } Q_2}\theta$$

$$\theta = \frac{\Sigma Count(A)\ \Sigma Count(B)}{\Sigma Count(A\cap B)\Sigma Count(C)}$$

$$A\Sigma Count(B) \triangleq \frac{\Sigma Count(B)}{\Sigma Count(\neg B)}$$

$$A\Sigma Count(B/A) = \frac{\Sigma Count(B/A)}{\Sigma Count(\neg B/A)}$$

$$\boxed{A\Sigma Count(C/A\cap B) = \frac{A\Sigma Count(C/A)A\Sigma Count(C/B)}{A\Sigma Count(\neg C)}}$$



SYLLOGISMS

p_1 ⟵ major premise

p_2 ⟵ minor premise

---

p ⟵ conclusion

disposition   proposition

$d_1 \xrightarrow{\text{restoration}} p_1(Q_1)$ ⟵ fuzzily quantified premise

$d_2 \xrightarrow{\text{restoration}} p_2(Q_2)$ ⟵ fuzzily quantified premise

$d \xleftarrow{\text{suppression}} p(Q)$ ⟵ fuzzily quantified conclusion

compositionality ⟶ $Q = f(Q_1, Q_2)$ independent of p_1 and p_2

most students are undergraduates

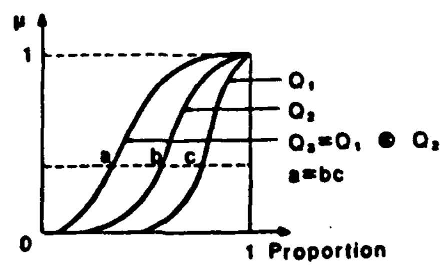
most undergraduates are young

---

most² students are young

INTERSECTION/PRODUCT SYLLOGISM

Q_1 A's are B's
Q_2 (A and B)'s are C's
$(Q_1 \otimes Q_2)$ A's are (B and C)'s
$\geq(Q_1 \otimes Q_2)$ A's are C's
$(Q_1 \otimes Q_2)$ A's are C's
monotonic



NEGATION SYLLOGISM

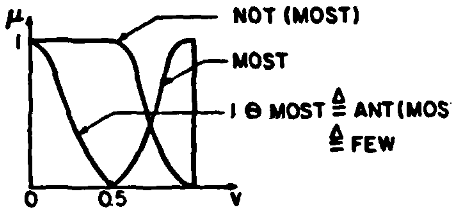
$$\frac{QA\text{'s are }B\text{'s}}{\geq(1 \ominus Q)A\text{'s are not }B\text{'s}}$$

$$\underbrace{\Sigma Count(\neg B/A)}_{is\ Q_1} \geq 1 - \underbrace{\Sigma Count(B/A)}_{is\ Q_2}$$

$$Q_1 \geq 1 \ominus Q_2$$

= if A is nonfuzzy

$$\frac{\text{few tall men are not fat}}{(1 \ominus \text{few}) \text{ tall men are fat}}$$



ANTECEDENT CONJUNCTION SYLLOGISM

$$\frac{\begin{array}{l} Q_1A\text{'s are }C\text{'s} \\ Q_2B\text{'s are }C\text{'s} \end{array}}{?Q(A \text{ and } B)\text{'s are }C\text{'s}}$$

$$\frac{\begin{array}{l} \Sigma Count(C/A) \text{ is } Q_1 \\ \Sigma Count(C/B) \text{ is } Q_2 \end{array}}{\Sigma Count(C/A \cap B) \text{ is } ?Q}$$

NO ASSUMPTIONS $Q = [0,1]$

MYCIN $Q = Q_1 + Q_2 - Q_1Q_2$

(Assumptions = ?)

PROSPECTOR: Assumption ≈

$$\boxed{\Sigma Count(A \cap B/C) = \Sigma Count(A/C)\ \Sigma Count(B/C)}$$

CONSEQUENT DISJUNCTION SYLLOGISM

$$\frac{\begin{array}{l} Q_1A\text{'s are }B\text{'s} \\ Q_2A\text{'s are }C\text{'s} \end{array}}{QA\text{'s are }(B \text{ or } C)\text{'s}}$$

$$\boxed{Q_1 \oplus Q_2 \leq Q \leq 1 \ominus (Q_1 \oplus Q_2)}$$

CONSEQUENT CONJUNCTION

$$0 \ominus (Q_1 + Q_2 - 1) \leq Q \leq Q_1 \ominus Q_2$$

DUAL:I r

NEGATION

$p \longrightarrow X$ is F

not $p \longrightarrow$ not (X is F)

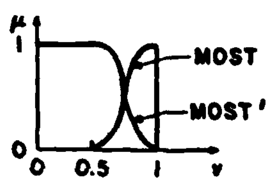
X is not F

QA's are B's $\rightarrow$ ΣCount (B/A) is Q

not (QA's are B's) $\longrightarrow$ ΣCount (B/A) is not Q

not (QA's are B's) $\longleftrightarrow$ (not Q) A's are B's

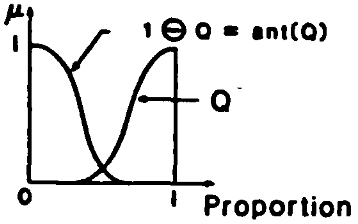
not (most tall men are fat) $\longleftrightarrow$ (not most) tall men are fat



NEGATION

$$\frac{QA\text{'s are }B\text{'s}}{(1 \ominus Q)\ A\text{'s are (not }B)\text{'s}}$$

$\underbrace{(1 \ominus Q)}_{ant(Q)}$



$$\frac{\dfrac{QA\text{'s are }B\text{'s}}{\geq (1 \ominus Q)A\text{'s are (not }B)\text{'s}}}{(1 \ominus Q)A\text{'s are (not }B)\text{'s}}$$

Q is monotone decreasing

CONSEQUENT CONJUNCTION SYLLOGISM

Q_1A's are B's

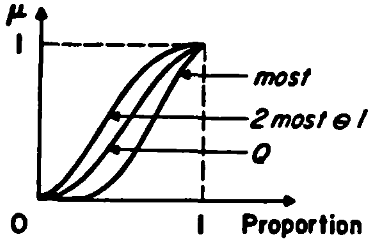
$\underline{Q_2A\text{'s are }C\text{'s}}$

?QA's are (B and C's)

$0 \circledvee (Q_1 \oplus Q_2 \ominus 1) \leq Q \leq Q \circledwedge Q_2$

$Q_1 = Q_2 = most$

$0 \circledwedge (2\ most \ominus 1) \leq Q \leq most$



CHAINING SYLLOGISMS

containment $(B \subset A)$

$$\frac{\begin{array}{l}Q_1A\text{'s are }B\text{'s}\\ Q_2B\text{'s are }C\text{'s}\end{array}}{(Q_1 \bullet Q_2)\ A\text{'s are }C\text{'s}}$$

$$\frac{\begin{array}{l}\text{most }A\text{'s are }B\text{'s}\\ \text{most }B\text{'s are }C\text{'s}\end{array}}{\text{most}^2\ A\text{'s are }C\text{'s}}$$

$$\frac{\begin{array}{l}\text{most students are undergraduates}\\ \text{most undergraduates are young}\end{array}}{\text{most}^2\ \text{students are young}}$$

CHAINING SYLLOGISMS

MAJOR PREMISE REVERSIBILITY

Q_1A's are B's ← major premise reversible

Q_2B's are C's

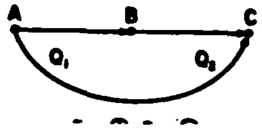
$\geq (0 \vee (Q_1 \oplus Q_2 \ominus 1))A$'s are C's

most American cars are big

most big cars are gas guzzlers

$(2\ most \ominus 1)$American cars are gas guzzlers

TRANSITIVITY OF CONTAINMENT



FUZZY SYLLOGISMS

Q_1A's are B's	most young women are slim
Q_2C's are D's	many slim women are attractive
?Q_3 E's are F's	?Q young women are attractive

Chaining: B = C, E = A, F = D

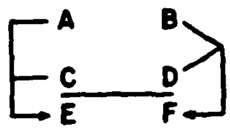
Consequent conjunction: A = C, E = A, F = B ∧ D

Antecedent conjunction: B = D, E = A ∧ C, F = B

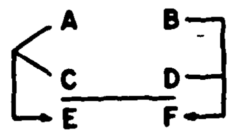
Intersection/product: C = A ∧ B, E = A, F = C ∧ D

Disjunctive syllogisms

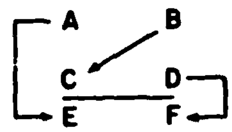
FUZZY SYLLOGISMS



consequent conjunction

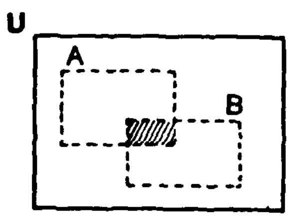


antecedent conjunction



chaining

PRODUCT RULE



Q_1U's are A's

Q_2A's are B's

$(Q_1 \otimes Q_2)$U's are (A and B)'s

$\geq(Q_1 \otimes Q_2)$U's are B's

$(Q_1 \otimes Q_2)$U's are B's (monotone Q or B ⊂ A)

most students are undergraduates

many undergraduates are freshmen

(most ⊗ many) students are freshmen

EXAMPLE

Intersection/product syllogism

Q_1 A's are B's

Q_2 (A and B)'s are C's

$(Q_1 \otimes Q_2)$ A's are (B and C)'s

almost_all A's are B's

all B's are C's

all (A and B)'s are C's

(almost_all⊗all) A's are (B and C)'s

(almost_all) A's are (B and C)'s

≥ almost_all A's are C's

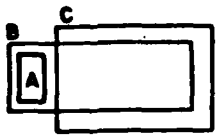
almost_all A's are C's (if monotonic)

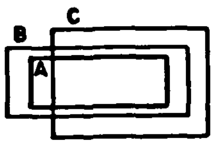
ROBUSTNESS

all A's are B's
almost all B's are C's

?QA's are C's

Q = none_to_all

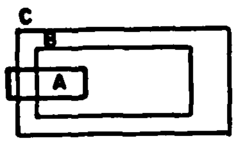




almost all A's are B's
all B's are C's

?QA's are C's

Q = almost_all



DEDUCTION

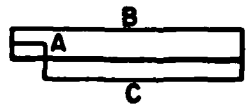
BRITTLE		ROBUST	
↓		↓	
p_1		p_1	$1-\epsilon_1$
p_2	≥ 0.99	p_2	$1-\epsilon_2$
p_3		p_3	$1-\epsilon_3$
•		•	•
p_n		p_n	$1-\epsilon_n$
q	$[0,1]$	q	$\geq(1-\Sigma_i\epsilon_i)$

In general, deduction in two-valued logic is britt

all A's are B's
almost_all B's are C's

---

none-to-all A's are C's



DISCUSSION ON PRESENTATION OF LOTFI ZADEH

DR. DeGROOT: Art, do you have any comments you would like to make?

DR. DEMPSTER: As the first discussant, I suppose it's my task to introduce the ferocious battles that Morrie was alluding to but I hope I will disappoint you on that.

I agree that in the real world there are many logics used and most of them are pretty fuzzy. In all different disciplines we reason in many different ways.

I certainly found everything that Lotfi said great and very appealing as heuristics and I think it's a fascinating field of study to think in terms of these heuristics, which is pretty much the way I think of what I've been listening to.

I might just mention this. In the side, since the words "representation" and "usually" and things came up, a colleague of mine, Fred Mosteller, wrote a long sequence of papers with Bill Kruskal in the Review of the International Statistical Institute on the concept of representation and what it means, and so on.

Fred has had a long interest in writing on sort of the heuristic side. I think Augustine (Kong) is involved in such a project at the present time, and he might be able to lead to some kinds of discussions you would be interested in.

My own need, however, is more in the direction of a need to defuzzify fuzzy logic. By that, I mean not taking the word "fuzzy" out of it since that's that the essence of it, but somehow clarifying the concepts of fuzziness.

Something that sticks in my mind is that R. A. Fisher, who is sometimes thought to be an obscure person, says someplace -- and I'm just paraphrasing it -- the wonderful thing about probability is that you can reason very precisely. The paradox is dealing with uncertainty in very precise terms and I think that's sort of what my focus is mainly on.

So I feel, as I'm sure Glenn does, a need to understand the mathematics. I think Glenn has made an advance in that in a recent technical report he goes over some of the concepts of marginalization and extension and combination and tries to relate the possibility measures to belief functions in various ways. I'm sure he will talk about that.

I suppose this can be debated. By being precise, of course, you get down to narrow models that are not going to stand up in the long run, but if you're not precise you sort of wind up going around and around the mulberry bush and not getting anywhere. So this is the kind of problem that troubles me.

The other thing that I think we need to do is to take some of the examples that Lotfi has brought up in recent writing and also in the draft paper (that he didn't follow precisely) and try to analyze them, not from my language or from his language, but from some language that we both understand as a common scientific approach to problems and see where that kind of things tries to lead.

Again I believe Glenn has discussed several of Lotfi's examples in his recent technical report and we need to do more of that in order to try to bridge the gap.

I enjoyed the presentation very much and some of the intuitive principles, like fuzziness getting worse the more things you try to combine, things of that sort which certainly have to be true but technically I think we need to get down to specifics of what the mathematics is and how to analyze examples in a common way.

DR. WATSON: I always enjoy hearing Lotfi speak because it makes me aware of how poor my understanding of the subject is when he manages to get so many different ideas into but a short frame of time.

I think at this time I'll just run through the Vu-graph I prepared and it addresses rather than the issues raised in his talk, this new concept of usuality which struck me as being a very interesting one which I need to go away and think about.

Rather than do that, I'll go back to some of the basics which some of you maybe wanted to ask about fuzzy set theory.

We've been clear that its a theory for theories imprecision rather than uncertainty and as Lotfi said, he sees it as being a companion to probability theory rather than a replacement for probability theory. I think that's an important point to make, so if we're using it in artificial intelligence systems we are using it in parallel with probability theory.

Of course, again one could back off from that and people have often asked this question and, fine, I can see that imprecision could be thought of as being a different concept from uncertainty, but if you're precise I'm uncertain and so it is always possible for imprecision that I find in the real world, for me to describe it personally in terms of uncertainty.

I think Lotfi mentioned this. He said he took it as a decision, an analytic decision, that he was not going to follow that route although he recognizes it's a route that could be followed.

I think it is important to recognize that that is a parting of the ways between the fuzzy and the Bayesian approaches to these things.

Now there are and always have been a whole lot of problems which people new to fuzzy set theory raise when looking at it. Lotfi has mentioned these already.

Where do the numbers come from? Well, in this context you can get back to the argument Glenn has been making a lot that what you need in talking about where the numbers come from is to compare something with canonical examples. In the Bayesian theory you've got the canonical example of balls and urn. In Shafer's theory you've got the canonical examples of the meaning of evidence and so on.

The fact that there aren't such in fuzzy set theory at present, (although I may be wrong there) is a cause of concern.

There is also the problem how you do actually represent fuzzy sets which stand for something. Lotfi put up some slides saying fuzzy sets meant usually. Well, I could presumably come up with another fuzzy set which looked similar which was also supposed to represent usually.

I don't have time to go into this. I suspect that the comeback from Lotfi would be it shouldn't matter just precisely what these fuzzy sets should look like, cause after all it's the theory of fuzziness.

Well, I don't know of any detailed studies which have looked at the implications of the output of the fuzzy analysis as a result of changing the input numbers or the shapes of the input numbers.

If the outputs are sensitive to these things, then it's crucial to know precisely where to get them from. If the outputs are insensitive I wonder whether the outputs say anything at all, but that's something we can come back to.

Another thing that is very often asked is why these particular connectives, why the max and the min. It's now well known that there is a whole host of connectives which could describe the connective, the and, and the or, which have the properties that you require, namely in the case of crispness they correspond to traditional logic but in the case of fuzziness they don't.

And there are, I know the existence of, though I haven't studied some studies which go into why one should use max and min. Maybe Lotfi could come back to that at some stage.

I don't think that is satisfactory. I've mentioned the interpretation there, but I think the last thing is something that is important and I think it does relate to artificial intelligence blandness. Lotfi has suggested that in fact this is a virtue of using some kind of fuzzy analysis.

The fact is that after awhile really the imprecision is so great that you can't really say anything, and I wonder whether that sort of blandness is something we actually do want in our artificial intelligence expert systems.

I suspect that the greater degree of representation of uncertainty which a Bayesian theory affords might be a virtue rather than a vice.

There are positive points I see in fuzzy set theory. As I've mentioned before, I don't hold with the view that the only way to handle uncertainty must be to use probability theory. As I said in my previous little speech, you can decide to not go along with the axioms that support subjective probability and refuse to place the bets, refuse to go in for Dutch books, and I don't see any reason why you shouldn't do that.

The positive points I see, therefore, as Art just said, you can think of it as a heuristic to support the way you think. but as a representation for natural language, it seems to me it has a lot going for it and that's what artificial intelligence systems are trying to do. As a tool for imprecise implication it seems to me it's a very sensible thing, but again I think of it as a heuristically reasonable thing to do rather than something which we must do out of some logic or necessity.

Perhaps the combination which Lotfi talks about, the representation of imprecision about probabilities, is the one which attracts me most. Well, I think I've said enough.

DR. DEGROOT: Thanks very much.

DR. ZADEH: First let me comment on a point that Professor Dempster made in his comments and that has to do with the desirability of having, let's say, some sort of rigorous mathematical foundation for the theory, or something that goes beyond talking about these things in a more or less qualitative fashion.

I do feel that there is such a need, but I also feel that there are limits to which one can aspire to attain that objective.

In other words, I think that as people become older, they begin to become more conscious of the limitations of what we can do physically or intellectually.

The same applies I think to science. Initially we have these grandiose ideas that we might be able to construct very simple models of the universe -- I think Einstein was driven by this sort of thing -- just one equation will explain everything, that we could have theories of probability that would tell you exactly what probability is and so forth.

Gradually I think, or sooner or later, it begins to dawn upon people that these may be unrealizable objectives that some of the questions, some of the issues and probabilities here that animated Bernoulli and LaPlace and people like that are still with us. They have not really been answered.

It's quite possible that we have to lower our sights and be satisfied with theories which do not answer all of these questions completely, and that is where the concept of a disposition again comes in the picture.

Dispositions, as I said, are assertions which are preponderantly but not universally true. Now mathematics is very allergic to dispositions. You cannot write a paper that would be accepted for publication in a respectable mathematical journal in which the conclusion would be in the form that usually something works or usually it's true. You cannot do that sort of thing. It has to be true, or it is not true. That's all there is to it.

Now if we adhere to that kind of a standard, then we are essentially shutting off ourselves from all sorts of human knowledge and we also make it possible to deal with expert systems because it's impossible, I think, to come up with a theory that completely and satisfactorily answers all of the conditions that arise.

In lowering the sights, however, you retreat as little as possible. I'm not suggesting that we retreat all the way to philosophy or something where you just wave your hands, but rather you then do what is done in fuzzy set theory, and that is you allow truths which are partial truths, you allow probabilities that are specified linguistically - likely, unlikely, very likely and so forth, you even allow these membership functions where the question of how do you find that value, .8 -- how do you do it rigorously, cannot be answered perhaps.

You live with this sort of thing. You say, well, there are limits to the process of precisiation and so long as I can come up with conclusions that in some sense make it easier for me to arrive at the decision, or diagnose the disease or understand natural language, or will do a number of other things, I'll be satisfied even though it may stop short of a complete explanation in the traditional spirit.

I think this is a point that has to be emphasized, that as I said in a number of places, when I wrote in fuzzy sets, it represents a retreat, and it represents also an attempt at finding an accommodation with the pervasive imprecision of the real world.

In other words, you're saying that the sights we set for ourselves, the goals we set for ourselves, are unattainable. You have to retreat a little bit.

Incidentally, let me just say one thing in response to a question about independence. I think Glenn answered the question in such a way that again may not be satisfactory in some sense to the people who expect a rigorous thing that is that we can define probabilistically. That's what Professor Dempster did in his paper and what Glenn used in his paper. In other words, you define independence, but when it comes to a practical situation if somebody asked you a question "are these bodies independent or not?" at that point we really don't have criteria which allow us to answer the question. I think this is what Glenn had in mind when he said you have to use your intuition, you have to use heuristics, and so forth. This is what it boils down to.

So the connection between theory and reality, that connection becomes a fuzzy one.

To go back to comments made by Stephen. As usual I think Steve was very succinct in his points.

I think that a useful application of fuzzy sets is in the characterization of probabilities. In other words, you take probability theory as it is, you don't modify it in any way, but you allow the probabilities to be fuzzy, which is a generalization of interval valued probabilities.

Now to say the probabilities are fuzzy is not the same as saying that you're dealing with second order probabilities. Many probabilists don't like second order probabilities.

I think when you say that something is likely, when you say things of this kind, you are really using a possibilistic characterization of what is basically a probability. In many situations our knowledge of probabilities is not really good enough to enable us to justify the use of numbers. We simply don't know that much about real world probabilities, and in fact if we put aside urns and cards and things of this kind that I used as canonical examples in texts on probabilistic theory, I think that most real world probabilities are not measurable.

The example that I use to illustrate that point is the following one. I used that example because I saw it in a textbook on operational analysis and they cited that as an example of an application of operations research type of approaches.

They said, well, suppose that you want to decide whether or not to insure your car, and so what do you do? Well, you have to take into consideration what's the probability that it might be stolen and various other things, and they assume they have numbers for various things, and they assume that you know there is a probability that the car might be stolen of .001 or some such number.

The question is where do you get that number from? My contention is that it is not possible to measure or find that number, that if somebody asks the question "what's the probability that your car might be stolen?" you cannot answer that question on any level, theoretically or empirically or any level whatsoever. The reason you cannot answer is that it is a unique sort of thing.

The information that is available about the theft of cars in the District of Columbia is not that relevant to the question of what's the probability that your car might be stolen, because it's a particular car.

So this is the old problem of unique things. In other words. you have much more information that the probability that you need is conditioned on all sorts of things, whereas the problem that you have is not conditioned to those things, and there is no connection between the two that can do you any good.

The problem is this -- and this is by no means an artificial example -- I think if you look at real world probabilities, you will find that most of them are like that. Most of them are not measurable, so that our perception that probabilities are well defined, they can be measured, is not realistic, not realistic at all.

The subjective point of view, where you relate probabilities to betting behavior in my judgment is also not satisfactory because it merely tries to explain one thing in terms of another thing which is just about equally ill-defined.

In closing then, let me say this, that there are problems having to do with the measurement of grades of membership which I however don't regard as a serious problem as some people do in the use of connectives, but all it boils down to is this; in effect it says that the real world is too complicated for simple theories. You cannot do that. You have to have a language which is sufficiently expressive to enable you to deal with imprecisely defined probabilities in situations in which "and" in one context has one meaning and in another context has another meaning; conjunction does not have really a fixed meaning, situations where implication does not have a fixed meaning, situations in which various predicates do not have fixed meanings, and so forth.

You have to accept that. You have to accept that and you have to lower your sights and be satisfied with conclusions which are not quite as precise as those that we expect of traditional theories. Thank you.

DR. DeGROOT: Are there comments from the floor, or questions?

DR. BROWNSTON: Lee Brownston from Carnegie-Mellon University.

I have a question which goes beyond fuzzy set theory and possibility theory and touches on the theory of belief functions as well as the nature of expert systems.

What are your ideas on whether these theories are normative versus descriptive? How do you validate them if you think that they are descriptive, and if they are normative how do you justify using one particular set of operators as opposed to another?

DR. ZADEH: Well, here's the situation. There are many rules of inference. For example, the composition rule for inference, and various other rules of inference.

These rules of inference can be tested in examples which are sufficiently simple to enable us to use our intuition. To the extent -- I usually test these things to satisfy myself that I'm not off on the wrong track.

Generally, and I haven't found exceptions to this thing, in the case of simple examples, canonical examples, they tend to agree with our intuitition, which then sort of encourages you to apply them to examples which are not so simple, to make it possible for us to use our intuition.

Here's the situation. In many cases these examples are such that you cannot use probability theory and things of this kind to come up with answers to those problems.

In other words, if I said something like usually X is F and usually something-something, and then I ask somebody-probability theory, okay, what can you tell me? A probabilist could undo it.

Here I disagree a little bit with Professor Lindley because Professor Lindley felt that probability theory, Bayesian theory can handle all of these problems.

My test then is to give a number of problems like that and say, okay, go ahead, solve, so we do not have here the comfort of being able to use some other theory and to be able to compare the results.

Now in general I tend to be somewhat wary of normative theory because I think that normative theories when they tend to disagree with human intuition, upon further inspection turn out to contain some assumptions that may not be warranted or some other things, so I tend to feel that if there is disagreement that the chances are it's the theory that's wrong rather than human intuition.

Of course there are cases where that is not so. Whatever I say is the disposition. In other words, that is usually the case but not always the case.

I take also some issue with Professor Tversky's examples. There is one example which suggests that people may assign higher probability to A than to B if A is a subset of B. He feels that this is counter-intuitive but to me it is not all counter-intuitive. In fact, I have transparency to show that. It's simply a matter of how you interpret these things.

People interpret the question in such a way that they answer in terms of probability of A given B even though the person that asks the question expects that the answer to be a probability of B given A, so the probability of B given A tends to be counter-intuitive, but if you interpret the answer as that to the question what's the probability of A given B, then it's perfectly intuitive.

So I tend to shy away from any pretense to normativeness.

Another thing which I have some reservations about is the principle of maximization of expected utility which is preconservative as the normative principle.

DR. DeGROOT: Nozer?

DR. SINGPURWALLA: First a general comment. I've heard the word "Bayesian" used here several times. I believe everybody who has used it has in mind the ordinary calculus of probability, just the way we've learned it. You're not referring to Bayesian inference per se, you're just referring to a use of the ordinary calculus of probability.

Now to the question. Professor Zadeh, I sensed a kind of inconsistency in one of your statements. You used the words "possibility of a probability" somewhere, and then later on you said that the notion of probability was not clear, or at least was not complete. You also said that the notion of subjective probability was imprecise. If that be the case, what did you have in mind when you said "possibility of a probability?"

DR. ZADEH: First let me respond to the first part of your comment.

I think that the term Bayesian is used in two different senses. The sense in which it's used by people who don't know too much about the probability theory, people in AI and so forth, when they say Bayesian they mean the application of the rules of probability theory.

DR. SINGPURWALLA: That's really what I was trying to emphasize.

DR. ZADEH: I think this is not the sense in which Professor Lindley would use the term Bayesian. There it has to do with ratio of subjective probability, what you do if you don't know probabilities, and so forth, the frequentist interpretation versus the Bayesian point of view, and so forth.

That gets into different issues. Nobody will question the use of the formula "probability of Y is the integral probability of Y given X." This is not the sort of thing we are talking of.

Then if I use the term Bayesian then it depends really who I'm talking to. If I'm talking to AI people I'm using it in this first sense. If I'm talking to people who are really probabilists then I'am using it in the second sense. That's the differentiation that one has to make.

Now with respect to the second point, could you just run over again what --

DR. SINGPURWALLA: You used the term "possibility of a probability." You also said the term probability was very unclear; so what exactly did you have in mind?

DR. ZADEH: Okay, here's the situation. Probability theory by itself is a very precise theory. The imprecision comes when you want to relate that theory to the real world. It's in the interpretation of symbols and various things that the difficulty arises.

This issue is usually avoided in texts on probability theory. In other words, if you read a typical book on probability theory there will be practically no discussion of subjective probability or things of this kind. This is an issue that's avoided.

Now in any theory, in any theory, you have that problem. It's the problem of correspondence. It's the semantics of the theory. This is really what it boils down to, and questions that Steve raised related to the semantics of this theory -- what do you mean by .8, what do you mean by this, what do you mean by that.

Now in the theory -- in fuzzy logic since probability and possibility are under the same roof, it's perfectly okay to raise the question "What is the possibility of probability," "What's the probability of possibility," and so forth.

So if I said that all I know is that a certain probability distribution lies in a certain set -- in other words, you have incomplete information -- for example, it's the set of normal distributions with certain variance, where the mean is between alpha and beta.

Thats a class of probability distributions. Now that class is the possibility distribution for probability distribution. You say "what are the possible probability distributions?"

Now as I said, in the case of possibility theory, possibility is a matter of degree, so if I said -- instead of saying it's normally distributed with the mean between alpha and beta, if I said that the mean value is close to five, that parameter is a fuzzy parameter and as a result of that, that possibility distribution will become -- it's a fuzzy sort of a thing so I am dealing with a possibility distribution of probability distributions with that possibility distribution being a matter of degree. This is what it means.

I think, and this is what I did in my 1979 paper, on fuzzy sets information where the Dempster-Shafer theory was generalized to situations in which the sets that you have are fuzzy sets and the basic probability numbers are fuzzy probabilities. That's the generalization, that was given in that paper.

DR. SOLAND: I'm Richard Soland from George Washington University, and very naively I would like to come back to the question of semantics because it seems to me that one of the benefits of the fuzzy set approach supposedly is keeping things in natural language, but I think that's perhaps a danger also in that people don't always understand the same things when they use natural languages. Sometimes it's cultural and sometimes it's individual.

I wondered to what extent this can have an effect on the operational nature of the theory.

Too often people, when they deal with somewhat quantitative problems in a semantic way, tend to be careless in being imprecise -- that is, not thinking clearly and carefully, and will perhaps say that usually it takes such-and-such an amount of time without thinking about that clearly enough to be precise, even in the sense of possibility and fuzzy set theory. I think in a lot of our analysis work we attempt to be quite quantitative in our modeling in order to put precision in, where lack of precision may cause errors in the analysis.

I wonder what kind of dangers might come into the analysis because of individual and cultural differences perhaps in implementing this theory.

DR. ZADEH: That's a really good point. I think that there is a great deal of misunderstanding there when it comes to the issue of meaning, understanding natural languages.

What is not sufficiently differentiated is the problem of understanding on the one hand and the problem of representation on the other hand.

The point of view that I take here is that the approach relates to representation of meaning, rather than to understanding of meaning. It's a language that allows you to represent a meaning so if you say something to me and I ask the question what do you mean by that -- I will not try to figure out what you mean. I will ask you the question what do you mean by that, and then this is the language that enables you to represent a meaning.

Now one of the transparencies showed and it's sort of related to the question here, what do you mean by usually exists and leads to the question that Prof. Dempster raised, so what I tried to do then is something like the following.

I ask the question what do you mean by F? For example, usually X is small. I say what do you mean by small. So you say small is this.

I ask you what do you mean by small. I don't try to try to guess. Then I ask what do you mean by usually. I say usually is this.

Now notice I allowed the usually to go to fuzzy so if your perception of usually is so poor that you cannot really draw a curve like that, you can draw something very fuzzy.

Now once you have explained to me what is meant by usually and what is meant by small, then I will take these two and I will go through the procedure which enables me to find what is the meaning of usually X is small.

So semantics basically is nothing more than the composition of the meaning of a complex entity from the meanings of its constituents. I'll supply the meaning of the whole thing. That's really what it boils down to.

Once the meaning of usually X is F is made more precise, this is the precisiation meaning, then I can reduce this thing to a problem with nonlinear programming or something like that.

Until then, I cannot do it because I really don't know what's meant by usually X is F and that is where classical probability theory will falter because classifical probability theory does not provide a language for the representation of the meaning of things like usually X is F. That's really what it does not do.

Let's take a simple problem. Suppose I say an urn contains a hundred balls, of which 40 are black and the rest are white. What's the probability that the ball picked at random is black, so okay, you divide one by the other and so forth.

But suppose I fuzzify the problem. Suppose I said that the urn contains approximately a hundred balls of which several are big. Instead of saying black and white I introduce something that is fuzzy, like size, or large.

What's the probability that the ball drawn at random is large? You'd be in trouble, because there is no mechanism for representing the meaning of several, large, approximately one hundred, and so forth. That's where the problem is going to arise.

DR. DeGROOT: Well, that stimulated the audience. Let me see those hands again and I'll pick one. Stephen, I'll give you another try.

DR. WATSON: Can I just come back briefly on that, Lotfi.

Supposing I give you what I mean by usually and what I mean by big, why should I go along with whatever calculations you do on those numbers since I don't see that there's a framework of necessity which makes it clear that those are the calculations I should do on these numbers.

DR. ZADEH: That relates again to the issue that was raised before at least tangentially and that's that within the system you have very few dogmas. In other words, there are default approaches. The default approach would be like the one that I've indicated. In other words, that's the standard approach.

However, if you want to interpret these things differently, if you want to combine them differently, if you want to instead of using maxi-min you wanted to use some of the T norms and so forth, it is perfectly allowable within the theory.

In other words, at any point you can override what are standard procedures in the theory and substitute something that in you judgment is a more accurate representation of what you expect of this sort of thing.

DR. WATSON: And how do I choose between such things if they give different answers?

DR. ZADEH: Here's the situation then. What you have to do is you have to make a study of these things. You have to have essentially a collection of these tools together with some comments, say this works, well, this situation and this has certain properties and this has certain properties and so forth, but if you have some idea as to what are the properties of these things then you pick the one that fits your perception best.

In the absence of that sort of a thing you just use the standard default rule that is within the system. An example of that would be the definition of connecting and there would be a standard rule there. If you don't like it, if you feel that, well, this doesn't really accord with what you have in mind then use such and such a rule.

That's why it is open-ended in some sense. In other words, you can substitute user-defined relations for whatever is stored in the system.

DR. DeGROOT: I think we could go on discussing this for much greater length but lunch is imminent.

I do point out that there will be more time for discussion. Keep your questions in mind. There's an hour set aside this afternoon from 4:00 to 5:00 for general discussion.

I want to thank all the speakers this morning and the discussants, and I want to commend Prof. Lindley for being so patient and keeping quiet. But he knows that he gets first crack this afternoon and I think that perhaps has something to do with it.

(Laughter.)

(Luncheon recess.)

THE PROBABILITY APPROACH TO THE TREATMENT OF UNCERTAINTY IN ARTIFICIAL INTELLIGENCE AND EXPERT SYSTEMS

by

Dennis V. Lindley
Somerset, England and George Washington University

Talk given at a conference on the calculus of uncertainty in artificial intelligence and expert systems, George Washington University, 27-28 December 1984. Supported by Grant DAAG29-84-K0160, U.S. Army Research Office, and Contract N00014-77-C-0263, Project NR042-372, Office of Naval Research, with The George Washington University.

1. INTRODUCTION

Our concern in this paper is not with a general discussion of artificial intelligence (AI) and expert systems (ES) but with one particular aspect of them, namely the occurrence of uncertainty statements within AI or ES. We discuss how they should be made, what they mean, and how they combine together.

Uncertainty is obviously present in most ES algorithms because experts can rarely be totally sure of the statements they make. Thus in medical ES, the presence of a symptom array does not invariably imply the presence of one disease, so that diagnosis is inherently uncertain. Even the symptom may exhibit uncertainty for doctors may differ in their interpretations (see section 10). Prognosis is clearly even more uncertain. When discussing purely deterministic procedures there may be some merit in introducing uncertainty. For example, chess is a game with perfect information yet AI programs sometimes incorporate uncertainty as a way of avoiding the terrible complexity of the game. So uncertainty, whilst perhaps not ubiquitous, frequently occurs. Our task is to study approaches to dealing with it within AI and ES.

2. THE INEVITABILITY OF PROBABILITY

Our thesis is simply stated: <u>the only satisfactory description of uncertainty is probability</u>. By this is meant that every uncertainity statement must be in the form of a probability; that several uncertainties must be combined using the rules of probability; and that the calculus of probabilities is adequate to handle <u>all</u> situations involving uncertainty. In particular, alternative descriptions of uncertainty are unnecessary.

These include the procedures of classical statistics; rules of combination such as Jeffrey's (1965); possibility statements in fuzzy logic, Zadeh (1983); use of upper and lower probabilities, Smith (1961), Fine (1973); and belief functions, Shafer (1976). We speak of "the inevitability of probability."

3. MATHEMATICAL AND PHYSICAL MEANINGS FOR PROBABILITY

Before defending the thesis, it had better be made clear what we mean by probability. Most emphatically, we do not just mean numbers lying between 0 and 1: it is more interesting than that. There are two ways of responding to a question about the meaning of probability. One is to describe the concept mathematically. The other is to consider its interpretation in the physical world. We consider both these responses.

Mathematically, probability is a function of two arguments: the event A about which you are uncertain, and your knowledge H when you make the uncertainty statement. We write $p(A|H)$; read, the probability of A , given H . The function obeys the three rules:

Convexity $0 \leq p(A|H) \leq 1$ and $p(A|H) = 1$ if H is known by you logically to imply A .

Addition $p(A_1 \cup A_2|H) = p(A_1|H) + p(A_2|H) - p(A_1 \cap A_2|H)$.

Multiplication $p(A_1 \cap A_2|H) = p(A_1|H)\, p(A_2|A_1 \cap H)$.

We could elaborate on these rules: for example, by discussing whether the events have to form a σ-field, whether the addition law holds for an enumerable infinity of events, whether $p(A|H) = 1$ only

if H is known by you logically to imply A , and in other ways. But these would merely add mathematical glosses to the key ideas that probability lies between 0 and 1 and obeys two distinct rules of combination. From these three rules, perhaps modified slightly, all the many, rich and wonderful results of the probability calculus follow. They may be described as the axioms of probability. We prefer not to describe them this way because, as will be seen below, they can be derived from other, more basic, axioms and consequently appear as theorems.

The interpretation of $p(A|H)$ is that it is your subjective belief in the truth of A were you to know that H was true. It is often referred to as subjective probability because it is ascribable to a subject, you; and also to distinguish it from another use of probability called frequentist or objective. This latter we shall call chance, thus avoiding the adjective for probability. It is convenient to think of $p(A|H)$ as a measurement: like a measurement of length or temperature. It measures belief, not temperature. Like all measurements it has a standard. We may take the simple example of balls in an urn. For you, $p(A|H) = a$ if you are indifferent between receiving a prize contingent on A , knowing H , and receiving the same prize contingent on a black ball being drawn at random from an urn containing a proportion a of black balls. Of course, other ways are possible. It is a defect of many other approaches to the measurement of uncertainty that they do not have a standard by which to judge their statements.

4. THE USE OF SCORING RULES

Having interpreted probability in two, important ways, let us turn to the defense of the thesis of the inevitability of probability. The task is to study uncertainty, particularly in the context of AI and ES. As scientists and engineers we would expect to measure our object of study , to describe the uncertainty numerically. If we agree to do this, we have to decide what rules the numbers obey: for example, can we add them, like lengths? One way is to think of possible rules and choose some that seem reasonable. This is the method of classical statistics, fuzzy logic and belief functions. There is another method.

Suppose that in expressing your belief in A , given H , you provide a numerical value a . In what sense is a a "good" measurement of your belief? De Finetti (1974/5) had the idea of introducing a score function, which scores your measurement or, as we usually prefer, your assessment of your uncertainty of A , given H . For two function f_0 and f_1 the score, when a is announced as the assessment, is defined to be:

$f_1(a)$ if both A and H are true,

$f_0(a)$ if H is true, but A false, and

zero if H is false.

De Finetti used the quadratic, or Brier score: $f_0(a) = a^2$, $f_1(a) = (1-a)^2$. With the quadratic, a near 1(0) will give a low score when A is true (false) and H true. If H is false the statement about A is irrelevant since it was made on the supposition of H .

Suppose now that you, or the expert in ES, does this with several event pairs (A_i, H_i) ; is scored on each and the scores added. Then de Finetti showed for the quadratic rule, that the values a_i must obey the rules of probability. Lindley (1982) generalized the result and showed that virtually any score leads to probability: some scores are eccentric and result in only two possible values for a whatever be A and H . A consequence of de Finetti's result is that someone using rules for the combination of the a_i that are not probabilistic--for example, those of belief functions--will have a worse score, whatever be the truth or falsity of the A's and H's , than the probabilist. Notice how eminently practical this approach is. The "expertize" of an expert could be assessed by keeping a check on his scores. Of two probabilists, either one may do better than the other, but both will do better than someone not using the probability calculus.

5. AXIOMATIC APPROACH

In an alternative approach we think about the concept of uncertainty and try to latch onto simple, basic principles that ought to be present in any study of uncertainty; such that any violation of a principle would, when exposed, make the argument look ridiculous. The principles, self-evident truths, are called axioms and from these we would hope to deduce, by mathematical reasoning, the rules that the numbers obey. Euclidean geometry is the famous example of this procedure when applied to the measurement of length. This programme was first carried out for beliefs in 1926 by Ramsey (1931). The best-

known example is Savage (1954). De Groot (1970) presents what is perhaps the most readable version. All these approaches lead to the result that the numbers must obey exactly the three rules of probability above. In other words, the 'axioms' of probability have been deduced from other, simpler ideas that more legitimately can, because of their self-evidentiary nature, be called axioms.

Let the converse be emphasized: any violation of the rules must correspond to some violation of the basic axioms, of those rules whose violation would look ridiculous. We really have no choice about the rules governing our measurement of uncertainty: they are dictated to us by the inexorable laws of logic. Of course, they are entirely dependent on the chosen axioms and the history of mathematics warns us not to be too complacent about the "sacred" rightness of axioms. But at the moment, the axioms are unassailled and all variants produce minor variants in probability.

6. COHERENCE

At this point we should perhaps digress to discuss an important aspect of the Ramsey/Savage/de Finetti approaches that is often overlooked. The discussion will also help to explain why non-probabilistic views have had some success in AI or ES even though the ideas are unsound. The rules of probability show how different uncertainty statements have to fit together. Thus, the multiplication rule above, refers to three assessments and says that one of them must be the product of the other two. Instead of "fitting together"

we talk of coherence. The results just described can be stated as showing that coherence can only be achieved by means of probability. We may say belief functions are incoherent (they do not obey the addition rule).

Coherence is not peculiar to the measurement of belief. It applies to all measurement: for example, of length. If ABC is a triangle with a right angle at B, it makes perfectly good sense to say $AB = 2$ or $AC = 4$ or $BC = 3$, or even to make two of these statements together. But make all three together and you are incoherent, for Pythagoras demands that $AC^2 = AB^2 + BC^2$, which is not true of the numbers given. Similarly one can say that $p(A_1|H) = 1/2$ or $p(A_2|A_1 \cap H) = 2/3$ or $p(A_1 \cap A_2|H) = 1/4$, but one cannot make all three statements simultaneously. The multiplication law replaces Pythagoras. It is curious that coherence is strictly adhered to with lengths but often ignored with beliefs, reflecting the immaturity of belief measurement.

And that explains why non-probabilistic procedures can sometimes appear sensible. The adherents never make enough statements for coherence to be tested. They only tell us the equivalent of $AB = 2$ and $AC = 4$, never discussing BC, for to do so would reveal the unsound nature of the argument.

7. BAYES THEOREM

One example of coherence is so important in AI and ES that we should perhaps consider it now. Interchanging A_1 and A_2 in the

above statement of the multiplication law and recognizing that $A_1 \cap A_2 = A_2 \cap A_1$, we immediately have that

$$p(A_1|H)p(A_2|A_1 \cap H) = p(A_2|H)p(A_1|A_2 \cap H) .$$

Using the equivalent result but with $\bar{A}_2$, replacing A_2, we have

$$\frac{p(A_2|A_1 \cap H)}{p(\bar{A}_2|A_1 \cap H)} = \frac{p(A_1|A_2 \cap H)}{p(A_1|\bar{A}_2 \cap H)} \frac{p(A_2|H)}{p(\bar{A}_2|H)} .$$

This is Bayes theorem in odds form. (The odds (on) A are simply the ratio t of $p(A)$ to $p(\bar{A})$: the odds against are the inverse of this. In practice they are usually quoted as t-1 on or t-1 against with $t \geqslant 1$). To appreciate what it says, temporarily omit H from the notation and language, recognizing that it is present in every conditioning event in the statement of the theorem. Then the result is that the odds, $p(A_2)/p(\bar{A}_2)$, of A_2 are changed, due to the additional knowledge of A_1, into $p(A_2|A_1)/p(\bar{A}_2|A_1)$ by multiplying by $p(A_1|A_2)/p(A_1|\bar{A}_2)$. The multiplier is called the likelihood ratio. It is the ratio of the probabilities of the additional knowledge A_1, given A_2 and then given $\bar{A}_2$. Thus an AI system faced with uncertainty about A_2 and experiencing A_1 has to update its uncertainty by considering how probable what it has experienced is, both on the supposition that A_2 is true, and that A_2 is false. Any other procedure is incoherent. Most intelligent behavior is simply obeying Bayes theorem. A high level of intelligence consists in recognizing a new pattern. This is not allowed for in Bayes theorem, nor in any other paradigm known to me. The simple AI systems that we have at the moment must be Bayesian.

8. A CHALLENGE

Let us summarize where we have got to in the argument. On the basis of simple, intuitive rules; or using a technique of scoring statements of uncertainty; it follows that probability is the only way of handling uncertainty. In particular other ways are unsound and essentially ad hoc in that they lack an axiomatic basis.

There is however more than just the inevitability of probability. There is the consideration that probability is totally adequate for all uncertain situations so far encountered. This is often denied. The following statements are taken from Zadeh (1983).

"A serious shortcoming of [probability-based] methods is that they are not capable of coming to grips with the pervasive fuzziness of information in the knowledge base, and, as a result, are mostly ad hoc in nature."

"The validity of [Bayes rule] is open to question since most of the information in the knowledge base of a typical expert system consists of a collection of fuzzy rather than nonfuzzy propositions."

Shafer (1982) says, in comparing belief functions and Bayesian methods, "The theory of belief functions offers an approach that better respects the realities and limitations of our knowledge and evidence."

I offer a challenge to these writers and to all who espouse non-probabilistic methods for the study of uncertainty: the challenge is that anything that can be done by these methods can be better done with probability. I think this is a fair challenge. It is a requirement that the method has been used and is not just a topic for

theorizing, which rules out some speculations in the alternative paradigms. If the challenge fails then we shall really have advanced: for an inadequacy in probability will have been exposed and the need for an alternative justified. The challenge is in the spirit of Popper who partly judges the merit of a theory on its capability of being destroyed; for the rich calculus of probability leads to many testable conclusions. It is also relevant to Popperian ideas because he has discussed certain inadequacies in probability. These have been disposed of by Jeffreys (1961).

As these words are being written it is impossible to know what challenges might arise. All that can be done is to take material already in the literature and examine that. I begin with fuzzy ideas.

9. PROBABILITY IN PLACE OF FUZZINESS

As an example of a fuzzy proposition Zadeh (1983) cites

"Berkeley's population is over 100,000"

He says it is fuzzy because "of an implicit understanding that *over 100,000* means *over 100,000 but not much over 100,000*" (his italics). (He might also have added that Berkeley is fuzzy. Does it refer to the town in Gloucestershire or that in California? And population: does it merely refer to permanent residents or are students included? These are not jibes: my point is that nearly all statements are imprecise.)

The probalilistic approach would be to give a probabilistic statement about a quantity *that can be evaluated*. The qualification is

is important, de Finetti has emphasized. As far as possible all probabilities should refer to propositions or events that can realistically be tested for truth or falsity. This is because we want to use them. It may be necessary to introduce other propositions but only as aids to the calculation of testable ones. (In statistics parameters are used for this purpose. An example in section 14 will use guilt of a suspect.) A possible quantity to discuss in the fuzzy statement is the answer the relevant city official in Berkeley would give when asked for the population of Berkeley. If this is denoted X , then the probalistic statement corresponding to that quoted is $p(X|H)$ where H is the knowledge possessed by the maker of the statement. It would have a mode a little over 100,000 if the statement is in H .

It is important notice that in applications it may not be necessary to specify the full probability distribution $p(X|H)$. For example, it may be enough to quote its mean, the expectation of X given H ; what de Finetti calls the prevision of X given H . More sophistication may require the variance of X , or equivalently, the prevision of X^2 given H . Fractiles of X are another possibility.

All fuzzy propositions of this type can be interpreted probabilistically in a manner similar to our treatment of Berkeley. "Henry is young" needs a little care. It clearly refers to Henry (whom I take to be a well-defined person) and an uncertain quantity X , his age. But the description is very vague. Made on campus, Henry might be only 19: made at a faculty dinner Henry might be 30: made in a home for senior citizens, he might be 65. Consequently

H is very relevant to this result. Without context $p(X|H)$ will need to be appreciable even for $X = 65$.

10. NUMERICAL EXPRESSION OF FUZZINESS

Another example is both more serious and more elaborate.

"John has duodenal ulcer (CF=0.3)"

(CF is an abbreviation for certainty factor.) It is a well-known feature of medical studies that many concepts are imprecisely defined and that a difficulty in using medical records resides in the varied use different doctors make of the same term. Nevertheless doctors find it useful to identify features like 'duodenal ulcer'. The situation can be described probabilistically by introducting Δ , an ill-defined but supposedly real ailment, duodenal ulcer, and also D_i the appreciation of duodenal ulcer by doctor i . The fuzziness of the concept can be captured by considering $p(D_i|\Delta)$ and $p(D_i|\overline{\Delta})$, the probability that doctor i will say John has duodenal ulcer both when John has, and does not have, true duodenal ulcer. (Useful comparison can be made with Bayes theorem above: Δ replaces A_2 , D_i replaces A_1 and H is omitted from the present notation.) Notice that Δ may not be a testable quantity. It is introduced as a parameter to facilitate the calculation of quantities that are testable. For example, if the above statement is made by a first doctor, what is the probability that a second will agree? $p(D_2|D_1)$ can be evaluated by extending the conversation to include Δ . For example, the D_i might be independent, given Δ .

This second fuzzy statement introduces a numerical measure in the form of a certainty factor, here 0.3. This contrasts with the apparently similar numerical assertion that the probability (on an undefined H) that John has a duodenal ulcer is 0.3 in at least two ways. First, CF's combine by rules that are different from those of the probability calculus, so that they would inevitability produce worse scores in an adequate test than would probabilities. Furthermore, these rules have no axiomatic basis and are merely inventions of fertile, unconstrained minds. The second difference between CF's and probabilities is that the operational meaning of the latter is clear whereas that of the former is not. We may say that probabilities have standards, possibilities do not. One standard for probability was mentioned above: balls in an urn. But expectation of benefit or a uniform distribution may replace these. All measurement requires a standard and CF's are dubious because they do not have them. What does CF = 0.3 mean?

The literature on fuzzy logic is vast, complicated and somewhat obscure. I have surely missed some examples that it would be useful to test against the challenge which remains: anything fuzzy logic can do, probability can do better.

11. INCOHERENCE AND BELIEF FUNCTIONS

We next turn from fuzzy logic to belief functions. I have already considered a good example of Shafer's (1982) in the discussion to that paper. It is repeated here partly because to do so is simpler for me than to take another one; and also because it is then possible to respond to Shafer's reaction to my probabilistic argument. Before

giving this it might be useful to exhibit incoherence in the use of belief functions. (The argument also applies to fuzzy methods.)

We follow Shafer and write BEL(A) for the belief in A, omitting reference to the conditioning event. Now it is possible that

$$\mathrm{BEL}(A) + \mathrm{BEL}(\bar{A}) < 1$$

(similarly for certainty factors). Write $\mathrm{BEL}(A) = a$, $\mathrm{BEL}(\bar{A}) = b$ so that $a + b < 1$. (Necessarily $a,b \geq 0$) Let us score such a belief using the quadratic rule. The possible scores are:

$$\begin{array}{ll} A \text{ true} & (a-1)^2 + b^2 \\ \bar{A} \text{ true} & a^2 + (b-1)^2 . \end{array}$$

Now replace a by a', b by b' where $a' = a + \varepsilon$, $b' = b + \varepsilon$ and $\varepsilon = \frac{1}{2}(1-a-b)$. It easily follows that $a' + b' = 1$ and that both

$$(a'-1)^2 + b'^2 < (a-1)^2 + b^2$$

and

$$a'^2 + (b'-1)^2 < a^2 + (b-1)^2 .$$

Consequently it is certain (irrespective of whether A or $\bar{A}$ is true) that beliefs a and b will score worse than probabilities a' and b', adding to one. The result generalizes with any score.

12. PROBABILITY IN PLACE OF BELIEF FUNCTIONS

Now for Shafer's example. Imagine a disorder called "ploxoma", which comprises two distinct "diseases": θ_1 = "virulent ploxoma", which is invariably fatal, and θ_2 = "ordinary ploxoma", which varies in severity and can be treated. Virulent ploxoma can be identified unequivocally at the time of a victim's death, but the only way to

distinguish between the two diseases in their early stages seems to be a blood test with three possible outcomes, labelled x_1, x_2 and x_3. The following evidence is available: (i) Blood tests of a large number of patients dying of virulent ploxoma showed the outcomes x_1, x_2 and x_3 occurring 20, 20 and 60 per cent of the time, respectively. (ii) A study of patients whose ploxoma had continued so long as to be almost certainly ordinary ploxoma showed outcome x_1 to occur 85 per cent of the time and outcomes x_2 and x_3 to occur 15 per cent of the time. (The study was made before methods for distinguishing between x_2 and x_3 were perfected.) There is some question whether the patients in the study represent a fair sample of the population of ordinary ploxoma victims, but experts feel fairly confident (say 75 per cent) that the criteria by which patients were selected for the study should not affect the distribution of test outcomes. (iii) It seems that most people who seek medical help for ploxoma are suffering from ordinary ploxoma. There have been no careful statistical studies, but physicians are convinced that only 5-15 per cent of ploxoma patients suffer from virulent ploxoma.

My reply was as follows. The first piece of evidence (i) establishes in the usual way that the chances for a person with virulent ploxoma to have blood-test results of types x_1, x_2 and x_3 are $0{\cdot}2$, $0{\cdot}2$ and $0{\cdot}6$. The second (ii) is subtler for two reasons: x_2 and x_3 are not distinguished in the data, and the patients in the study are not judged exchangeable with other patients so that the chances β in the study and γ for the new patients are not necessarily equal. The first presents no difficulty since the likelihood for the

data is $\beta_1^r(\beta_2+\beta_3)^{n-r}$ where $r = 0{\cdot}85n$ and n is the number of patients in the study. The distribution of β given the data can therefore be found. Let $p(\gamma|\beta)$ be the conditional distribution of γ, given β. This concept replaces the single figure of 75 per cent quoted by Shafer and which yields a discount rate of $\alpha = 0.25$. It would be possible to suppose $\gamma = \beta$ with probability $0{\cdot}75$ and is otherwise uniform in the unit interval in imitation of belief functions; but this may be an unrealistic description of the situation. The third piece of evidence (iii) says the distribution of the chance θ that a patient has virulent ploxoma, $p(\theta)$, is essentially confined to the range $(0{\cdot}05.\ 0{\cdot}15)$. We are now ready to perform the requisite probability calculations.

Let G be the event that a new patient, George, has virulent ploxoma and let g_i be the result of his blood test. We require $p(G|g_i,E)$ where E is the evidence. From (iii) $p(G) = \int\theta p(\theta)d\theta$. From (i) $p(g_1|G,E) = 0{\cdot}2$ for $i = 1,2$ and $0{\cdot}6$ for $i = 3$. From (ii)

$$\begin{aligned} p(g_i|\bar{G},E) &= \int\int\gamma_i p(\gamma|\beta)p(\beta|E)d\beta d\gamma \\ &= \int E(\gamma_i|\beta)p(\beta|E)d\beta \end{aligned}$$

and the calculations can be completed in the usual way using Bayes' theorem. If $E(\theta) = 0{\cdot}10$, $E(\gamma_i|\beta) = \beta_i$ and $E(\beta_2|\beta_1) = \frac{1}{2}(1-\beta_1)$ then the probabilities of G given g_i are respectively $0{\cdot}025$, $0{\cdot}229$ and $0{\cdot}471$.

It may be objected that this analysis virtually ignores the uncertainty about the study and about θ. It does so because they are irrelevant. The interested reader may like to consider the case of

George and Henry and their blood tests. Then the uncertainties will matter: for example, $E(Y_i^2|\beta)$, involving the conditional variance of Y_i, will arise.

Shafer in response says that "Lindley insists that the uncertainties affecting this study are irrelevant and should be ignored. Is this reasonable? Suppose that instead of having only 75% confidence in the study we have much less confidence. Is there not some point where even Lindley would chuck out the study and revert to the prior 5-15%?" My reply is that Shafer is correct and that the uncertainty does matter a little, for it affects $E(Y|\beta)$. Were we to have no confidence at all in the study then $E(Y|\beta)$ would not depend on β, and $p(g_i|\bar{G},E)$ would be simply $E(Y_i)$ about which no information is given. (The prior on θ seems irrelevant).

Consequently I feel that the challenge has been well met with the example and, by a Popperian argument, the credibility of probability theory is increased.

13. COMPLEXITY, COVERAGE, DECISIONS AND RICHNESS

Here are four miscellaneous remarks.

(1) It should be noted that fuzzy logic and belief functions are considerably more complicated concepts than those of probability. With belief functions we start effectively with probabilities over the power set of the original events, itself much more complicated than the original set, and then have to elaborate on that. Dempster's rule of combination is vastly more involved than Bayes and then only applies in certain cases. Fuzzy logic leads to non-linear programming and

contains great complexities of language and ideas. Yet probability is extremely simple, using only three rules and containing rich concepts like independence and expectation.

Certainly if my challenge fails it will be necessary to introduce some change into probability ideas, which will almost surely increase the complexity, yet be necessary and rewarding. But until that happens is it not best to accept the advice of William of Ockham and not multiply entities beyond necessity?

(2) It is not implied in the challenge that probability can handle every problem involving uncertainty: the claim is merely that probability can do better than the alternatives. I believe that it has the potentiality to solve every uncertain situation but there are some for which the available techniques are inadequate. It is absurd to think that any paradigm can quickly resolve every relevant puzzle; some may resist solution for decades. For example, the medical problem of handling large numbers of indicants in diagnosis is currently unresolved because we do not have adequate techniques for handling the complicated dependencies that exist. (And certainly belief functions do not.) We need more research into applied probability and less into fancy alternatives.

(3) Why do we want to study uncertainty? Aside from the intellectual pleasure it can provide, there is only one answer: to be able to make decisions in the face of uncertainty. Studies that do not have the potentiality for practical use in decision-making are seriously inadequate. An axiomatic treatment of decision-making shows (Savage (1954),

De Groot (1970)) that maximization of expected utility is the only satisfactory procedure. This uses, in the expectation calculation, the probabilities and these, and only these, are exactly the quantities need for coherent decision-making by a single decision-maker. Only the utilities, dependent on the consequences, not on the uncertainties, need to be added to make a rational choice of action. How can one use fuzzy logic or belief functions to decide? Indeed, consider a case where $BEL(A) + BEL(\overline{A}) < 1$. Because you have so little belief in either outcome do you, like Buradin's ass, starve to death in your indecision between A and its negation? Reality demands probability.

(4) It is sometimes said, as in the quotes from Zadeh above, that probability is inadequate. This sense of inadequacy sometimes arises because people only think of probability as a value between 0 and 1, forgetting the whole concept of coherence and, in particular, ignoring the addition and multiplication laws. In fact probability is a rich and subtle concept capable of dealing with beautifully delicate and important problems. This richness is hard to convey without deep immersion in the topic. In order to display this, and also to try to avoid the impression that this paper is entirely concerned with bashing other ideas, I conclude by discussing a situation that arises in forensic science or criminalistics. It has been much discussed in the literature; a convenient reference is Eggleston (1983). An almost identical problem has been considered by Diaconis and Zabell (1982) using Jeffrey's rule. For reasons given below, I think their treatment is unsatisfactory.

14. A PROBABILITY EXAMPLE

A crime has been committed by a person who is to be found amongst a population of $(n+1)$ people. One of these is referred to as the suspect, the others are labelled in a non-informative way from 1 to n. Let G_s be the event that the suspect is guilty, G_i that person i is $(1 \leq i \leq n)$. Initially $p(G_s) = \pi$, $p(G_i) = (1-\pi)/n$ for all i. (Some forms of the problem have $\pi = (n+1)^{-1}$, which probabilistically does not distinguish the suspect from the other n.)

An investigator studying the crime says "the evidence suggests the criminal is left-handed." This is a fuzzy statement and its probabilistic interpretation requires care. After discussion the investigator says that the probability that the criminal is left-handed is P. This is still ambiguous. Diaconis and Zabell appear to interpret it to mean that the probability that the criminal will be found amongst the left-handers in the group of $(n+1)$ is P. I think a British forensic scientist would mean that if he had the criminal in front of him, the probability that he would be found to be left-handed is P. The former is the chance of guilt amongst left-handers: the latter of left-handedness amongst the guilty. Also the former requires reference to the population: the latter does not. Typical forensic evidence makes no mention of a population, only of the criminal, and so the latter interpretation is appropriate. There is a confusion between $p(A|B)$ and $p(B|A)$.

Working with the forensic interpretation, the formal statement is $p(\ell_i|G_i) = P$, where ℓ_i denotes the event that person i is left-

and

$$p(G_i|\ell_s\ell_1) \propto p^2(1-\pi)/n \quad \text{for} \quad 2 \leqslant i \leqslant n .$$

Thus $p(G_s|\ell_s\ell_1) = P\pi/\{P\pi + P(1-\pi)/n + p(1-\pi)(n-1)/n\}$. (2)

Rearranging the denominator as $P\pi + p(1-\pi) + (P-p)(1-\pi)/n$ we see that (2) is less than (1): the knowledge of another left-handed in the population has slightly decreased the probability that S is guilty. Notice that when $n = 1$, $p(G_s|\ell_s\ell_1) = \pi$: the evidence that all the population is left-handed has not changed the suspect's probability for guilt at all.

Evidence E_3 . There are no left-handers amongst the n people.

Combined with E_1 this means that the suspect is the only left-hander. Denoting E_3 by ℓ_0 , a use of Bayes theorem similar to that employed with E_1 and E_2 gives

$$p(G_s|\ell_s\ell_0) \propto p(\ell_s\ell_0|G_s)p(G_s) = P(1-p)^n\pi$$

and

$$p(G_i|\ell_s\ell_0) \propto p(\ell_s\ell_0|G_i)p(G_i) = p(1-p)^{n-1}(1-P)(1-\pi)/n .$$

Hence

$$p(G_s|\ell_s\ell_0) = P\pi/\{P\pi + p(1-\pi)(1-P)/(1-p)\} . \quad (3)$$

This clearly exceeds $p(G_s|\ell_s)$, equation (1), if $P > p$, showing that E_3 increases the probability that the suspect is guilty. Indeed, if $P = 1$, (3) gives 1 as it should.

Evidence E_4 . There is at least one left-hander amongst the n people.

E_4 is the negation of E_3 and may be written $\bar{\ell}_0$. It differs from E_2 in that the latter names a specific left-hander, #1. We have

$$p(\ell_s\bar{\ell}_0|G_s) = p(\ell_s|G_s) - p(\ell_s\ell_0|G_s) = P - P(1-p)^n$$

handed $(1 \leq i \leq n$ and $i=S)$. It was emphasized in the discussion of Bayes theorem that it is essential to consider the evidence A_1 both on A_2 and on $\overline{A}_2$. So here we need, in addition to $p(\ell_i|G_i)$, $p(\ell_i|\overline{G}_i)$. The latter is the chance that anyone is left-handed and may ordinarily be equated to the frequency of left-handedness in the population, p say. So $p(\ell_i|\overline{G}_i) = p$ for all i, including S. Presumably $P > p$. (In some forms of the problem $P=1$ and the forensic evidence is firm. This can realistically arise when dealing with blood types that can be identified without error.) Diaconis and Zabell do not consider p. This seems strange because the presence of an unusual trait intuitively carries more weight than a common one. The formal analysis below will confirm this.

15. THE ROLE OF ADDITIONAL EVIDENCE

Now consider various forms of additional evidence.

Evidence E_1. The suspect is found to be left-handed. In the notation this is the event ℓ_s. Simple use of Bayes theorem

$$p(G_s|\ell_s) = p(\ell_s|G_s)p(G_s)/p(\ell_s)$$

yields

$$p(G_s|\ell_s) = P\pi/\{P\pi + p(1-\pi)\} \tag{1}$$

which clearly exceeds π. E_1 is indicative of the suspect's guilt.

Evidence E_2. Person #1 is left-handed. This is ℓ_1. Now with both E_1 and E_2

$$p(G_s|\ell_s\ell_1) \propto p(\ell_s\ell_1|G)p(G) = Pp\pi .$$

Similarly

$$p(G_1|\ell_s\ell_1) \propto Pp(1-\pi)/n$$

and

$$p(\ell_s\bar{\ell}_0|G_i) = p(\ell_s|G_i) - p(\ell_s\ell_0|G_i) = p - p(1-p)^{n-1}(1-P) .$$

A further use of Bayes theorem gives

$$p(G_s|\ell_s\bar{\ell}_0) = \frac{P\pi - P(1-p)^n\pi}{P\pi + p(1-\pi) - (1-p)^n\{P\pi + p(1-\pi)(1-P)/(1-p)\}} . \qquad (4)$$

If $n = 1$ this give π in agreement with $p(G_s|\ell_s\ell_1)$, equation (2). It is easy to see that $p(G_s|\ell_s\bar{\ell}_0) < p(G_s|\ell_s)$, equation (1), so that E_4 slightly decreases the probability of the suspect's guilt.

Now for a subtlety: compare (2) and (4), that is the probability that the suspect is guilty given, in (2), the name of a left-hander and in (4) the mere presence of a left-hander. These are different. It is not too hard to verify by induction on n that

$$p(G_s|\ell_s\ell_1) < p(G_s|\ell_s\bar{\ell}_0)$$

for $n > 1$, so that the definitive knowledge of #1's left-handedness reduces the suspect's guilt probability by more than does the mere evidence of someone's left-handedness.

I leave the reader to think out whether the following argument is correct. Knowing there is a left-hander in the $n(E_4)$, no information about the suspect's guilt can possibly be provided by telling me the number of one of them. Accepting this, you are told it is #1. Since (2) and (4) differ (and calling #1 Smith for dramatic effect) the evidences "Smith is left-handed" and "There are left-handers, one of whom is called Smith" have different evidential value.

16. CONCLUSION

Our argument may be summarized by saying that probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate. The justification for the position rests on the formal, axiomatic argument that leads to the inevitability of probability as a theorem and also on the pragmatic verification that probability does work. My challenge that anything that can be done with fuzzy logic, belief functions, upper and lower probabilities, or any other alternative to probability, can better be done with probability, remains.

REFERENCES

De Finetti, B. (1974/5). Theory of Probability (2 vols.). New York: Wiley.

DeGroot, M. H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.

Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability. J. Amer. Statist. Ass. 77, 822-830.

Eggleston, R. (1983). Evidence, Proof and Probability. London: Weidenfeld and Nicolson.

Fine, T. L. (1973). Theories of Probability: An Examination of Foundations. New York: Academic Press.

Jeffrey, R. (1965). The Logic of Decision. New York: McGraw-Hill.

Jeffreys, H. (1961). Theory of Probability. Oxford: Clarendon Press.

Lindley, D. V. (1982). Scoring rules and the inevitability of probability. Int. Statist. Rev. 50, 1-26 (with discussion).

Ramsey, F. P. (1931). Truth and probability (In The Foundations of Mathematics and Other Essays. London: Kegan, Paul, Trench, Trubner, 156-198.

Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton: University Press.

Shafer, G. (1982). Belief functions and parametric models. J. Roy. Statist. Soc. 44, 322-352 (with discussion).

Smith, C. A. B. (1961). Consistency in statistical inference and decision. J. Roy. Statist. Soc. B 23, 1-37 (with discussion).

Zadeh, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. Fuzzy Sets and Systems 11, 199-227.

TRANSCRIPT OF ORAL PRESENTATION BY DENNIS LINDLEY:
PROBABILITY CALCULUS
FOR THE TREATMENT OF UNCERTAINTY

DR. LINDLEY: My thesis this afternoon is extremely simply stated, that the only satisfactory description of uncertainty is probability, that if you do it in any other way then in some sense it will be defective.

We had better start, I think, by getting clear what I mean by "probability." There are two ways of answering the question "what is probability?"

The first answer is within the mathematical framework, you can say what is the mathematics of the subject. The second way to answer the question is to say what it means in the world.

Let me take both approaches. The notation that I'm going to use is P of A vertical line H to mean the probability of an event A, given information H. (Slide 1)

The first point I want to make is that probability is a function of two things: the event A about which you are uncertain and the information H that you have when you make your statement of uncertainty.

There is a lot of nonsense talked about probabilities as a function of one argument. That is clearly nonsense because if your information changes, obviously your uncertainty about the situation can change and so consequently your measure of uncertainty will depend on your information as well as on the event whose uncertainty you're considering.

So we have a function of two arguments and though I'm sure everybody in this room knows them, I have just written down the three basic laws of probability.

The first one, convexity, says that probability lies between zero and one, which of course is all that most lay people know about probability, and also that the probability of A given H is equal to one if you know that H logically implies A. That is down there to make sure you can tell the difference between truth and falsity.

The next law is the addition law that says that the probability of either A_1 or A_2 occuring is the probability of A_1 plus the probability of A_2 minus the probability that they both occur. One is usually looking at that when A_1 and A_2 are exclusive and so this last event cannot occur, and then we have straightforward addition.

Finally there is the multiplication law, the probability that the two events both occur is the probability of one of them times the probability of the second given that the first is part of your information.

Notice that the multiplication law is the only law in which the information changes. So it plays a very central role in probability discussion. One could easily have omitted H from the first two laws but not the third.

Now many people think of those as axioms, the axioms of probability. One of the points I want to stress this afternoon is that to me they're not axioms at all, they are theorems. In fact, one of the most beautiful pieces of modern mathematics that I know is De Finetti's proof of the multiplication law.

That's the mathematical answer. If you ask me what is probability, I say mathematically it's anything that obeys those laws. The next question is what does it mean?

Now I want to stress this following point. It does seem to me to be tremendously important and yet other people somehow don't seem to think it is.

We are trying here to measure something. We are trying to measure uncertainty. Now if you want to measure anything in this world, you have to have a standard of reference.

For example, if I wish to measure this desk in yards, I have to do it essentially with reference to a standard.

Several years ago there would have been at the National Bureau of Standards a standard yard. There was a standard yard in Britain. There was a standard meter in Paris. All measurements were referenced to that standard.

If I want to measure temperatures there's a standard, zero referring to freezing of water, et cetera.

Every measurement that you make is with respect to a standard and here we are trying to measure uncertainty and so what I want to know is what is our standard.

Well, you can have several standards. You can have a standard yarc at the National Bureau of Standards, or you can have a standard based, I believe, on the wavelength of sodium light. There are lots of standards.

Here is a standard. We have an urn in which there is a proportion a of black balls. If now with respect to any event A and information H you say that the probability of A given H is a, then the standard is the following: that you are indifferent between a prize contingent on A, and the same prize contingent on a black ball being drawn at random from the urn.

You are going to get a prize either if A occurs or if you get a black ball and if you are indifferent between those two, then that is your probability.

Now that is the interpretation of probability and one of the criticisms that I have to make of other points of view, for example, fuzzy logic and belief functions, is that they do not have a standard, or if they do I can't understand what it is. They do not have a standard by which to judge things.

So there we have probability both mathematically and interpretatively. Now my thesis, remember, is that the only way to measure uncertainty is by means of probability.

Let us now take what I think is a really rather practical point of view. Let us imagine that we are going to have a series of people and they're going to measure uncertainty in any way they choose. They have agreed they are going to use numbers. I don't care how these numbers come or how they use them as long as they have some method of doing it.

Now let us imagine we watch these people do this. They all assign their uncertainties for various things and we ask a simple question. "How good are they at doing it?"

For example, suppose we were trying to measure lengths of tables. You know each person could measure the length and put the answers together. In some way we would get somebody down from the National Bureau of Standards who is really super at measuring lengths and he'd measure and we would compare them and see how good they were. A very simple little problem.

How are we going to do that with uncertainty? If somebody says the uncertainty is .8, is that good or is it bad?

Well, it clearly depends on whether the event is subsequently seen to be true or not.

If, for example, we agree that the bigger the number the more likely the event is to be true in some sense, a .8 when the event is true is somehow better than .2 when the event is true.

On the other hand, if the event is false, .2 is better than .8.

Now De Finetti had, I think, a brilliant idea that what we could do is score people. So let me now introduce you to the idea of a score function. (Slide 2)

Let the uncertainty of A given H be described by a number a. I don't care how you've got a, you can do it by fuzzy logic, you could do it by belief functions, you can do it by sampling theory statistics, you could do it by Jeffrey's rule, you can do it in any way you like.

All I'm saying is let's suppose that you were to assess the uncertainty of A given H by little a, and now we're going to score you.

If A and H are both true, you will get a score which depends on A, a function of A. Let's put suffix 1 there corresponding to A being true.

If on the other hand A is false, you will get a score f zero of A.

If it turns out that H is false, there is no score at all because your assessment of uncertainty was conditional on H, so if H is true you're not in the game.

Now let us suppose then that these people, however they get these numbers little a, are scored. What we're going to do is keep a tally on their scores and add up all the scores.

Now that seems to me a very practical way of doing this sort of thing and in fact I understand that it's done in meteorology.

People make a statement about the uncertainty of rain tomorrow and wait and see whether it rains tomorrow and give them a score. This is repeated over several days and the scores added. A good meteorologist gets a low score and a bad meteorologist gets a high score. (I'm thinking of these scores as penalty scores. They are bad things. You want to minimize them. You can turn yourself upside down if you like and make them good, but my convention is going to be that.)

The simplest score function is the quadratic score function, sometimes called the Brier score function, f_1 of a is (a minus 1) squared and f_0 is a squared.

Suppose the event A is a sunny day and you give it value .8. If the event then turns out to be true, your score is going to be .8 minus 1, that's .2, all squared, a little score of .04.

On the other hand, if the event is false, you're going to get .8 squared, you're going to get .64, you're going to get a big penalty score, you've done rather badly.

So .8 has done rather well if A is true, and done rather badly if false.

Then we're going to take all these scores and we're going to add them up. Now that seems to me a very sensible system of doing these things. You know, I'd like to take these columnists who are making forecasts and these other people in different fields making forecasts, and just check them. I'd love to take some sampling theory statisticians and just see how well they do with their inferences.

Now De Finetti proves a most remarkable result. He proved that with this quadratic score function, those numbers a had better obey the rules of probability. Whatever happens, whether the A 's there are true or false, you will do better if you make those numbers obey the rules of probability. Those are the three rules that I had on my first slide.

And so consequently De Finetti proved the rules of probability. There were theorems resulting from these assumptions.

VOICE: What does it mean, "do better?"

DR. LINDLEY: The score will be less, whatever happens.

VOICE: Is it the expectation of the score ___

DR. LINDLEY: No, it's for sure. There is no expectation involved in this. It's for sure. Whether the events are true or false for sure is here.

Now you might say, well, that's an interesting result but I think you've sort of cheated because you have made this score go near to one for truth and zero for falsity. I've really forced it into being probability, haven't I? But lo and behold I've not because it turns out almost, whatever those two functions f_1 and f_0 are the same result persists.

Whatever function you take there the numbers that you get will obey the laws of probability, at least with a little catch.

Here's the catch. Suppose, for example, these two are exponentials. Not quadratic, but exponentials. The numbers that you would turn out to be giving would be the logarithm of the odds rather than the probabilities, so in other words I would have to turn all those probability rules into logodds rules, which could be done. They'd look just a bit messy in log-log form, but that's what would happen.

What happens is if you take almost any score function, the person will give you a known transform of probabilities. What transform it is depends on f_1 and f_0.

Now there are some strange score functions that don't do this. There are some strange score functions that would lead you always to give one of two answers, say zero or one. There are some score functions that push you into giving one of two numbers, always zero or always one, never anything else, but those are very strange score functions and one doesn't want to use those. It's like making everyone say true or false in response to a question. That would be silly.

So consequently, if you were to use any reasonable score function then the numbers that you would get would, possibly after a transform, obey the rules of probability.

Now the key point here is this. The key point is that the rules by which these numbers manipulate are not arbitrary. You can't sit down and think up some clever rules. Let me quote somebody who said something this morning. Somebody said about MYCIN: they made up their own calculus. Well, all I can say that if they made up their own calculus they were silly, because they're not at liberty to make up their own calculus. You can't say, "Oh, I rather like the supremum or I rather like the infimum". You can't do it. The rules must be the rules of probability.

Of course I made some assumptions in deriving the result of the inevitability of probability: essentially that the (arbitrary) scores added. Surely a modest assumption: how else would you combine the scores?

Now there is another approach that gives the same answer and one must just mention it. It's usually called the axiomatic approach and in this country the famous originator of it is L. J. Savage. Here you put down some reasonable axioms and you deduce from those axioms the rules of probability again.

The best exposition of that that I know is in our chairman's book, Optimal Statistical Decisions, in which the axioms are beautifully spelled out and the argument goes through and he proves that the numbers have to obey the rules of probability.

Let me summarize this by saying, to me probability is inevitable. This is the inevitability of probability. There are no other ways of doing this job except in terms of probability. Any other method will surely produce a larger penalty score. This is not a matter of expectation. The argument is based on surely doing this.

Now you might say, well, if it's like that, if that is the situation, why have people been doing these funny things, why don't they use it?

Well, one of the reasons is that people don't always make enough statements for their stupidity to be revealed. Let me give you an example of this. (Slide 3)

Suppose that you were to say the probability of A_1 given H is a half and then you were to say that the probability of A_2, given A_1 and H, is 2/3.

Now those two numbers could be anything you like, any numbers between zero and one. The Bayesian world is a very free world. You can have any numbers that you like there, but once you have chosen those two numbers, your freedom has completely and utterly disappeared if you now think about the probability of the intersection of A_1 and A_2. It must be a third.

Now clearly I'm not going to trap anybody if all they will give me are the first two statements. They can be any numbers.

I'm only going to be able to trap them when the third comes in as well, so therefore I have to be rather forceful in this scoring business. I have to demand of people that they make logically related statements.

Let me give you an analogy which is not quite perfect but might help drive it home.

Suppose that we were doing measurements, ordinary Euclidian geometry, and we were going to talk about right-angle triangles. Each one of you in this room told me about the lengths of the two sides of the right angle; you told me the height and the base.

Everyone in this room could give me a pair of numbers and provided they were positive I couldn't query them. But as soon as any one of you gave me the third side of that triangle I would be onto you like a shot, because Pythagoras' theorem would tell me what that third number would be, and anybody in this room that gave me a number that didn't satisfy Pythagoras' theorem, you would all say, oh, he's crazy.

That's the same situation here. The first two statements can be any old numbers, but... The reason, it seems to me, that many people in their arguments don't fall into the difficulty is because they don't allow themselves to go near the difficulty, they don't give themselves the chance of exhibiting it.

It's very easy to say that this is .3 and this is .8 and this is .7, but if you combine those things together then you get yourself into difficulty. This is called "coherence."

Now I felt that I really just had to say something about Bayes' theorem. The subject is rather peculiarly called Bayesian statistics. (Slide 3)

I've written out Bayes theorem there in its simplest form. I have omitted H from the notation, so that you've got to add an H all the way through. It just says that the odds prior to A transform into the odds posterior to A by multiplying by the likelihood ratio.

The most beautiful example of this that I know is in a court of law. Let the event A be that the defendant is guilty. On the right are the odds on him being guilty before A , which is some evidence, comes along and on the left are the odds after the evidence. It says that what you have to do is to multiply the odds before you get the evidence by the ratio of probability of the evidence on the assumption that he was guilty to the probability of the evidence on the assumption that he was innocent.

Those are the two things that are relevant. At the end of this talk I will give you an example of somebody only using the numerator there, only trying to get through with the numerator and forgetting the denominator. Of course that will produce a curious and unsatisfactory answer.

I now am going to make a challenge. The challenge is this: that in the study of uncertainty anything that can be done by whatever it is, can be better done by means of probability. (Slide 3)

This is my challenge to you. I make it not in any arrogant or conceited way. I think this is the way that science proceeds. Science proceeds by somebody setting up a theory, setting up a coconut shy, and trying to destroy it.

Those of you who are familiar with the work of the British philosopher, Karl Popper, will know that's the keystone of his argument. The argument is that what a scientific theory should first do is to have lots of deductions that can be made from it.

There is nothing to be said for a theory that says each planet has got behind it an angel pushing it around. There's nothing in that theory because it doesn't tell you where the planets are going, but if you take a theory of Newtonian attraction you can work out where the planets are going to be, and deduce lots of things.

Having deduced all these things, you then test them and see if they're right and you try to destroy the theory, and as long as you can't destroy it you enhance the theory.

So I'm giving you enormous opportunities to destroy this theory. I challenge you; that anything you can do by fuzzy logic, anything you can do by belief functions, can be better done by probability. Notice that is the caveat there: anything that can be done by fuzzy logic. I'm not saying that probability could do anything. I think it probably can, but I'm not saying that. What I'm saying is if it can be done that way it can be done by probability and it can be done better that way.

I can't respond to that challenge immediately because I don't know what you're going to say so what I've done is I've gone through the literature a little and taken some examples and discussed them.

Here's an example from fuzzy logic. This example is taken I believe from one of Zadeh's papers. (Slide 4)

The statement he quotes is Berkeley's population is over 100,000. He says that this is a fuzzy statement, because over 100,000 is fuzzy -- it means a little bit over 100,000 but not too much. I would agree with him, it is fuzzy. What he doesn't also say is the rest of the thing is fuzzy as well.

What does he mean by population? Does he include the students who are only there for part of the year, or is it only the residents?

Berkeley. Where is Berkeley? Is he referring to the town in Gloucestershire, England, or the one in California?

Everything is fuzzy. Every statement is fuzzy. There's nothing peculiar about the over 100,000.

Now the probability approach to this would first of all say, we've got to think about something well defined. Let me make an assumption that he was talking about Berkeley, California, which is part of the information H, of course. Let us assume that part of H includes the knowledge it was Berkeley, California.

Now what we would agree to do, I think, might be to go down to the relevant official -- I don't know what he's called in the United States -- in Berkeley and ask him what is the population. That will be a number and therefore we could make a probability statement about X, and then we could go find out what X is, and we could score it. It will have, of course, to be done on the basis of whatever information is available.

So often people forget the information. It may happen that one of you in this room actually knows the population of Berkeley, in a sense or you may know what the official figure was last year, or something like that.

So here is a statement in fuzzy logic that can perfectly well be turned into a probability statement.

Now let's take another one, a little more sophisticated. John has duodenal ulcer with CF equal .3. CF is certainty factor of .3. (Slide 4)

That one is a little more sophisticated (and a little more serious where it's dealing with someone who is sick for one thing) but the real thing is it's got a number attached to it of .3.

Now that is certainly very fuzzy. Let's assume John is a definite person, so that there is no fuzziness about him. The really interesting thing about that is the duodenal ulcer, if I understand it correctly, is not very well defined. Consequently, what one has to do is to think about a concept of duodenal ulcer without being really clear what it is.

On the other hand, there are statements by doctors that John has got a duodenal ulcer and those are firm statements. He did say duodenal ulcer, so consequently the probabilistic approach would introduce two things. We tend to use the Greek alphabet for things that we can't actually get in touch with directly, and the Roman alphabet for things that we can. Delta would correspond to this vague thing, duodenal ulcer. D_i would be the i-th doctor's statement that he has duodenal ulcer.

The sort of thing we're interested in is if the doctor said he's got duodenal ulcer, (that's part of our information) what is the probability that he truly has duodenal ulcer?

Now notice that this statement has associated with it a number. Well, there are two queries. What is the standard? It's not the question of how did he get .3, but what does it mean. How can I check that value of .3? Balls and urns, or whatever it is.

Another point I want to make is those .3's cannot be combined by the rules of fuzzy logic because if you do combine them using the supremum and infimum you will do worse for sure if De Finetti were to come along and score you. You'll just tote up the scores and see that the fuzzy person had a larger penalty score than the probabilist.

Let me now turn to belief functions. One of the properties of belief functions is that the belief in event A -- I'm omitting H here for ease of notation - the belief in A, plus the belief in not A can add up to something less than one. So if I denote belief of A by little a and the belief in not A by b, a plus b can add up to something less than one. (Slide 5)

Well, now let's score a person. I say that a person who does this will for sure score less than a probabilist who makes them add up to one.

a and b are adding up to something less than one, so the total deficiency is one minus a minus b, so let us take that deficiency and add half of it to a and the other half to b giving me new numbers a prime and b prime.

So if I started with .4 and .2 adding up to .6, the deficiency is .4, half of it is .2 and I'm now going to add .2 to each of them.

It's very easy to show that the total scores will be less with a prime and b prime. On the the fourth line is the score when the event is true. On the fifth line is the score when the event is false.

Whether the event is true or the event is false, the probabilist with his a prime and b prime will for sure get the smaller score than the belief function person.

In my paper, I had dealt with one of Glenn Shafer's beautiful examples concerned with an imaginary medical disease called ploxoma which I discussed before. There isn't the opportunity to discuss it with you now in any sort of detail.

Let me just extract from the argument one point. Let us suppose that we have some data, very obvious and straightforward data, in which a number n of trials have been carried out and r of those have resulted in success. We have r successes out of n trials and let us suppose that the chance of success on any trial is beta.

I have to just say here that I am using the word chance in a different sense from which Glenn Shafer seemed to be using it this morning which is why I asked him that question. We can discuss this if need be.

But there is a chance beta, and the likelihood is the familiar binomial likelihood, $\beta^r (1-\beta)^{n-r}$. There is the situation.

Now in this example of Glenn Shafer's he says, well it may be that those trials have some relevance to what you're interested in, but they're not quite the same.

For example, let us suppose that those trials are being carried out in medical patients in England and here we are in the United States. Well, we may well say to ourselves the chance beta in England is different from that in the United States.

On the other hand, the British study does tell us something. It's not entirely irrelevant. We feel that if we were to do the same sort of study in Washington we wouldn't have exactly the same thing, but we'd have something like it. The two diseases or whatever it is that we're studying are perhaps not quite the same thing, but they're similar.

This is a very real situation and I thought a very fine point to bring out, the fact that quite often one has data that is of some relevance to what you're studying but doesn't fit absolutely perfectly. So you cannot say there's that data with chance beta, here I have another situation with chance beta and so now I learned something about beta from those r statistics out of n trials, and I can now apply it here. That's not true in many cases.

Well, what do we do? The simplest thing to do is to imagine that in Washington the chance is some number gamma. Gamma is not the same as beta but they're related. What we would do is say we're uncertain about gamma and so we would think about the probability distribution of gamma given beta, which is an assessment of how like the Washington population is to the relevant British population in respect of what this success is that we're talking about. Then one can infer what the probability of gamma is by the usual rules of probability.

There is one point, incidentally, that comes out of this argument, that when you have to study these problems, in a sense you don't have to think. By that I mean is that when you're doing this discussion you know very well that all you have available are those three rules of probability and nothing else. Everything follows from the rules of probability.

Consequently, if I want to get hold of gamma, I know very well I've got to go and use the rules of probability and that the relevant law of probability will be there. It's a recipe. It's a rule for carrying out the calculation.

Consequently, there is no need, in my view, for any belief function structure connecting the population in England and the population in Washington. You can do it perfectly well by probabilistic arguments.

Now let me make four rather miscellaneous points. Complication: Professor Zadeh (I think I quote him correctly) said this morning the real world is too complicated for simple theories. (Slide 6)

I couldn't disagree more. I was brought up in a small town in England near the little village of Ockham and in the 13th century there lived in this village of Ockham a gentleman named William, and William of Ockham said that entities should not be multiplied beyond necessity.

I suppose I learned this when I was young but it still seems to me to be extremely good, an extremely valid principle that you should make the situation as simple as you possibly can and that seems to be admirably met by probability.

There are only three rules of probability, whereas the number of entities knocking around in fuzzy logic grows by the hour. Possibilities. Now today, what is it we have today? Usualities. The complexity grows and grows and grows. I don't think that's the right way to go.

The right way is the way that William of Ockham said to us. Make the situation as simple as you can. If it doesn't work, Okay, you'll have to make it a little more complicated, and if it doesn't work again make it a bit more complicated.

In fact, if you do make it complicated, you are almost certainly wrong.

I don't know too much about modern theoretical physics, but when I talk to theoretical physicists for a moment they are very worried because their models are getting too complicated. Everybody is looking around for that simple thing because they believe, as Einstein did, in simplicity. It looks as though Hawking in Cambridge is getting very near to it, there is a simple rule underlying it all.

Simplicity is a thing much to be admired. I'll always remember the time I spent at the Harvard Business School with Bob Schlaifer. Schlaifer said to me one day. "people love to delight in complexity, it obscures all their mistakes."

I think there is a lot to be said for that. Complexity is to be abhorred. It's not the right way to think about things. Simplicity is.

The next question I'd like to discuss is a tricky one of coverage. My challenge is that anything that can be done by these other methods can be done by probability. The challenge is not that everything can be done by probability, but that may well be true.

I can give you problems that I as a probabilist cannot solve. Let me give you a very, very simple one indeed, that occurs in a simplified form of medical diagnosis where there are a number of symptoms, each of which is either absent or present. There are 40 of them.

The probability structure of 40 symptoms, each of which are 0-1 variables, is extremely complicated and I certainly do not know how to handle it myself for the moment but that's not to say that probability is the wrong approach.

It is well known that when you take a scientific theory, if it's good a theory it poses lots of very difficult problems. The usual example quoted is the three-body problem in Newtonian mechanics. When Newtonian mechanics was first formulated, people thought it would be very easy to study the motion of three bodies and it turned out to be a very difficult technical problem.

The fact is that there are lots of problems about uncertainty that we probabilists cannot solve at the moment. I think that they are mostly technical difficulties.

Now another point that was made this morning: decision making. The axiomatic approach that I mentioned of Jimmy Savage's leads to the rules for decision making and the rules for thinking about uncertainty are probabilistic. The rule for making decisions is the rule of expected utility.

What I do not understand is how we are supposed to make decisions on the basis of belief functions or fuzzy logic.

You remember that slide I had up: (Slide 5) let's put it back again. It can happen that the belief in A plus the belief in not A is less than one. I don't see how you're going to use that sort of situation.

You know the story, do you, of Buradin's ass? Buradin's ass was placed equidistant from two equally succulent bundles of hay and he starved to death because he could see no reason for preferring one bundle over the other one.

Belief function people seem to be in the exactly same situation. They put .2 here and .4 there, leaving the other .4 that's over. What are they going to do with that .4; starve to death? They have to make up their minds. They have to act. At least real people have to act, and in order to act, it seems to me that you have got to use the full force of the probability argument.

So my question there is, how on earth can these arguments be used in decision making?

It is sometimes argued that probability theory is inadequate. I have pursued some of these arguments and it doesn't seem to me that it is inadequate. In fact, it seems to me to be quite the contrary.

If you carry through the probability argument, I at any rate am continually surprised with the richness of the results that it produces.

Let me give you a very simple example. This occurs in a paper by Diaconis and Zabell in the Journal of the American Statistical Association a couple of years ago. It is concerned with a trial, a criminal, and some evidence about the criminal being lefthanded.

Now when you do the probability calculations, it turns out that one piece of probability that you have to put in is the probability that a person taken at random from the population is left handed. This is not entered at all into the Diaconis and Zabell argument.

Now if I go through the probability mechanism, it turns out that I have got to think about the probability of a random person from the population being lefthanded and I say to myself, well, is that right or is it wrong, and surely it is right and it is a relevant thing.

If lefthandedness is very rare, the evidence that the person is lefthanded says much more than if lefthandedness is a very common thing.

What is happening is that I carry through the probability argument and I find that certain things enter into the situation and I say to myself, "well, is that right?" and it always happens in my experience that it is right, that those things are indeed relevant.

I think we had an example this morning, though I'm not quite sure, when Glenn produced his example with icy conditions and the thermometer, and he had to bring in a P and a Q and it seemed to me, thinking about it very quickly, that it was very right that P and Q should have entered into the argument and if they did not, then the argument surely is unsatisfactory.

Surely the arguments about lefthandedness is unsatisfactory if it doesn't take account of the rarity of lefthandedness.

I have just a few minutes left and I would like to conclude with a little example from probability that may interest you. The full example will appear in the paper and I put it in order partly to be constructive. I don't want to appear to be knocking everybody down (which, of course, I am) but I wanted to appear to be a little bit constructive, and I wanted to try and show you an example of what seems to me to be the extreme sublety of the probability argument. You might say when you've seen this that it isn't subtle at all, but it came upon me somewhat as a surprise and I think it's come upon other people as a surprise.

Here is the little problem. The problem is this; it's being discussed in the legal literature quite a lot.

A crime has been committed and it is known that the criminal lies in a set of n+1 people, one of whom is a suspect and there are n other people, so there is a suspect S and n other people and they are numbered 1, 2, 3...up to n, and the numbers contain no information. Of course in reality they would be Smith and Jones, et cetera.

The event that we've interested in, of course, is whether the suspect is guilty. That is the event G suffix S. (Slide 6)

Then there are the other events, that the other persons are guilty. G is the event that person number i is guilty.

Now the evidence is produced that the criminal is lefthanded with a value of .8. Now the first thing you have to ask yourself is what does that mean.

In the paper from which I've taken it, it is held to mean that if we took the population of lefthanded people -- that is, if we took all the lefthanders in here, the probability is .8 that we would find the criminal amongst them. That is, conditional on lefthanded, the probability of the criminal being there is .8.

I don't think that's what I mean and I'm quite certain it's not what a British forensic scientist would mean. A British forensic scientist would mean that having got the *criminal in front of him*, his probability of his finding him to be lefthanded would be .8. I say British scientists because I have no experience with what American ones would do in this context.

The British forensic scientist would say had I got the criminal in front of me, there is a probability of .8 he would be lefthanded. That is a statement where the event A you're uncertain about is lefthandedness and you're given that he's the criminal.

The other statement, the one I had before, is a statement in which lefthandedness is in H and A is being the criminal so they're upside down statements.

I've written the two statements out. (Slide 6). The first statement, the criminal would be found amongst the lefthanders, and the second statement, given the criminal the probability of being lefthanded is .8.

I'm going to use the latter interpretation. It is the one that seems to me to be right.

Now some evidence comes in. The first piece of evidence is that the suspect is lefthanded. Now as soon as that comes in you begin to think he might be guilty, because you've already had the information that the criminal has high probability of being lefthanded.

Notice how the rarity of being lefthanded comes in. Lefthandedness occurs in only about 10 percent of the population, so the probability of being lefthanded for an ordinary person is about .1.

This evidence comes in. Now another piece of evidence comes in. The other piece of evidence is that person number one is also lefthanded. Now that sends the probability of guilt down a little, doesn't it, because there's another lefthanded person knocking around and so the probability that the suspect is guilty is going down.

Now imagine another piece of evidence. This piece of evidence is that there is a lefthander amongst those n. You're not told that number one is the lefthander. You're told that there is a lefthander.

I've written that as the negation of 1_0. 1_0 means that there aren't any lefthanders. The information, the evidence there, is that there is a lefthander amongst the n.

Let me just recapitulate. There is the evidence that the suspect is lefthanded for sure. No fuzziness or anything about that. He is lefthanded. There is the evidence that number one for sure is lefthanded. There is the evidence that there is somewhere amongst those n people a lefthanded person.

Now we can calculate the probability the suspect is guilty given the evidence that he is lefthanded and person number one is lefthanded.

You can also calculate the probability that the suspect is guilty given that the suspect is lefthanded and that there exists another lefthanded person and those two are different.

That was my surprise to me and the sublety of the argument to me. These two are not the same. In fact, the former one is always less than the latter.

So now we have a curious situation. Someone has told you that there is a lefthanded person amongst numbers one up to n, and this is the probability that the suspect is guilty.

Now suppose I say, oh, yes, there's a lefthanded person amongst that group of people, it won't do any harm, will it, if you tell me his number.

They can't give you any information to tell what the number is so consequently if you now say, oh, his number is one, you appear to be in the first situation and that probability is not equal to that one.

That seems to me a pretty subtle and curious state of affairs and to describe probability theory as being inadequate in a situation like that does seem to me to be rather strange.

It's a beautiful and it's a rich and it's a wonderful subject, and I commend to your attention that probability is the only satisfactory measure for uncertainty.

(Applause.)

1

$p(A|H)$ probability of event A, given information H.

Convexity $0 \leq p(A|H) \leq 1$ and $p(A|H) = 1$ if H is known by you logically to imply A

Addition $p(A_1 \cup A_2 | H) = p(A_1|H) + p(A_2|H) - p(A_1 \cap A_2 | H)$

Multiplication $p(A_1 \cap A_2 | H) = p(A_1|H)\, p(A_2|A_1 \cap H)$

STANDARD: Urn with proportion a of black balls. $p(A|H) = a$ if indifferent between prize contingent on A, given H, and ball at random being black.

2

SCORING RULE. Let uncertainty of A given H be described by number a.

Score $f_1(a)$ if A, H both true

$f_0(a)$ if A false, H true

0 if H false

E.g $f_1(a) = (a-1)^2$ $f_0(a) = a^2$

Scores for a_i for (A_i, H_i) $i = 1, 2 \ldots$ added.

Then scores make a's obey laws of probability.

AXIOMATIC APPROACH

INEVITABILITY OF PROBABILITY

Any other method will surely produce a larger (penalty) score.

3

COHERENCE

If $p(A_1|H) = 1/2$ and $p(A_2|A_1 \cap H) = 1/3$ then $p(A_1 \cap A_2 | H) = 1/6$

BAYES THEOREM

$$\frac{p(A_2|A_1)}{p(\bar{A}_2|A_1)} = \frac{p(A_1|A_2)}{p(A_1|\bar{A}_2)} \; \frac{p(A_2)}{p(\bar{A}_2)}$$

odds posterior to A_1 = likelihood ratio $\times$ odds prior to A_1

CHALLENGE. In the study of uncertainty — anything that can be done by .... can be better done by probability.

4

"Berkeley's population is over 100,000"

X = population H = information

$p(X|H)$

"John has duodenal ulcer (CF = 0.3)"

D true duodenal ulcer

D_i doctor$_i$'s statement of duodenal ulcer

$p(D|D_i)$

What is standard for CF = 0.3

Cannot combine by fuzzy rules

5

$$\mathrm{BEL}(A) + \mathrm{BEL}(\bar{A}) < 1$$

$$a + b < 1$$

Define $a' = a + \frac{1}{2}(1-a-b)$ $\quad b' = b + \frac{1}{2}(1-a-b)$

$(a'-1)^2 + (b')^2 < (a-1)^2 + (b)^2$ and

$(a')^2 + (b'-1)^2 < (a)^2 + (b-1)^2$

DATA r successes in n trials : chance β

likelihood $\beta^r(1-\beta)^{n-r}$

A similar, but not identical set of trials : γ

$$p(\gamma|\beta)$$

Then $p(\gamma) = \int p(\gamma|\beta)\, p(\beta|\,\text{data})\, d\beta$

6

COMPLICATION Occam's razor

COVERAGE Technical difficulties

DECISION-MAKING. Expected utility

INADEQUACY.

S suspect , n other persons

G_S guilty , G_i $(i=1,\dots n)$

Criminal is left-handed : 0.8

" Criminal found amongst left-handers "

" Given criminal, prob. of l.h. is 0.8 "

Suspect is left-handed L_S

#1 is left-handed L_1

$\exists$ l.h. amongst n $\bar{L}_0$

$$p(G_S|L_S L_1) < p(G_S|L_S \bar{L}_0)$$

DISCUSSION ON PRESENTATION OF DENNIS LINDLEY

DR. DeGROOT: I'm sure that Glenn Shafer and Mr. Zadeh would like to have a chance to take up the challenges that were thrown out, but perhaps we should have our discussants comment first.

I think we should rule out allowing anyone to say the same thing at the end of more than three of the four talks. You can only say the same thing twice.

Art, do you have some comments?

DR. DEMPSTER: Sure. I think I agree with almost everything Dennis says and I think the way he says it is wonderful. The agreement I guess is modulo the usual amounts of fuzziness on both our parts. Well, I probably didn't agree with his first sentence, either.

I don't know that this counts as repeating myself having said this morning we all need to be exposed more to ideas of probability. I think it's marvelous that we've had this proselytizing talk.

My own experience was that I read Feller Volume I in the early 1950s and I've been convinced about probability ever since then, so that things like Savage's axioms, and scoring rule theorems, and Dutchbook arguments I think are all very pretty but I was already convinced, so they didn't interest me a great deal.

I do think that there's a slight misrepresentation here about belief functions. Belief functions are based on the theory of probability so almost everything that Dennis is saying really is helping support belief functions if you look at it in the right way.

I would like to make one comment that I think is at a slightly deeper level. One thing that I've learned, or at least learned to say, from reading a little recently about artificial intelligence, glancing only an hour or so at David Marr's book, is this notion that things can only be understood if you look at them at many different levels.

My reaction to what Dennis is telling us is that he's giving us a perfect story at one level but it's too closed in. It isn't relating far enough out into the world.

Another thing that the AI people preach or tell us which is very valid and something I've been saying about statistics for a long time is that you can't understand it unless you relate it to the goals, the problems, the thing that's being worked on.

I might comment, perhaps not so specifically on what Dennis is saying but a comment of the chairman this morning that he couldn't see why we wanted to specify two numbers rather than one, it's hard enough to do one so to try to do two it's at least twice as hard, or probably much more from Dennis' point of view.

The thing is that there's one of these levels that's going on that's concerned with constructing these probability models in relation to the goal of the probability analysis and that's something that Glenn has written a great deal about, something that I think is the essence of the problem, not what Dennis is talking about since I've believed in that for more than 30 years.

The essence of the problem is how to we construct these things in a way that has some kind of scientific validity and it just does seem to happen that when you do that (for example in Glenn's example about Fred) you come up with probabilities, sure, but they get reflected in ways that lead to upper and lower probabilities or beliefs and plausibilities, whatever kind of terminology you want to use.

That I think is operating at a whole different level of understanding that the theory that Dennis has talked about doesn't seem to relate to.

DR. WATSON: I think it's very difficult to know how to respond to Dennis' very clear argument of the inevitability of probability, but it has to be faced because some people in this room don't share his conception of this inevitability and you then have to ask what's wrong, is it the argument or is it the premises.

I think there are two premises to the argument which are worth looking at, and I won't say much about them but it's something that other people may have thought of.

Firstly, is judging probability the same as judging length? You'll notice that part of the argument is predicated on the assumption that it really is the same sort of thing.

I think it's not, but I don't think we ought to spend time talking about that now.

The second point is that the argument he presented from De Finetti was based on scoring rules and I was reminded of a little verse that was on the wall of a house I stayed in when I was young, about the Great Scorer of Life. It went, "And when the one Great Scorer comes to write against your name, he writes not if you won or not but how you played the game."

(Laughter.)

Now that version of Victorian morality is probably not terribly appropriate for this afternoon's discussion but it did strike me that to go along with the argument that Dennis is making, you had to presume that this scoring mechanism was a sensible one and I'm suggesting that one may refuse to go along with that part of the argument and this of course allows you an out from the conclusions of the theory.

The two points I want to take are in different order so I'll have to show them both at the same time, I'm afraid. Fuzzy set theory, you could argue, is not concerned with uncertainty. It therefore does not claim to be a contender for these numbers that are supposed to represent uncertainty. It's describing some different human perception.

Now you may say that you can't understand what perception it is it's describing, but in my view, it is a perfectly valid position to take that it's worth thinking of, there being a different perception to uncertainty, namely imprecision and that fuzzy set theory may be an appropriate thing for describing imprecision.

I'd like to finish with this last point that Dennis made so forcibly about the sense of adopting Ockham's Razor, which I share. I think we all share.

I would argue that conversely the application of probability actually leads to enormous complexity and that what we need is a theory which leads to simpler representations than is provided by probability theory.

David Schum of Rice University has done quite a bit of work on the application of probability reasoning in legal contexts. He's done a very nice paper which analyzes the famous legal case, that of Salmon's pills, that of whether Salmon's pills killed somebody or not. I advise you to go and read the papers if you want a good analysis of a very simple legal inference.

Now he applied a form of Bayesian thinking, of probabilistic thinking, to that case and concluded that the number of probability judgments that one needed to make in order to come to some sensible conclusion was beyond all reason.

Not only that, but there was not one but a large number of different ways one might set about structuring the problem.

I therefore think that if you're going to adopt Ockham's Razor, and I do, you might be led elsewhere than to probability theory. Thank you.

DR. DeGROOT: Dennis, do you want to respond to those comments before we move on?

DR. LINDLEY: No.

DR. DeGROOT: Glenn, do you have comments that you would like to make, or do you want to wait and see how the discussion goes?

DR. ZADEH: I have a comment. I think that one cannot really take issue with the conclusions that Dennis drew from the particular example that he considered involving scoring, but I think that one can question the big jump from whatever conclusions that one can draw from that example to the much more sweeping statement concerning the inevitability of probability theory. I think there is a big gap there.

I hated to get into it because I have to figure out if what I'm going to say next is a repetition of what I've said already. (Laughter.)

In the first place, even in that example the assumption is that something is either true or false, but what happens if your forecast is such that it's a matter of degree?

For example, the forecaster says it will be a rainy day and it is rainy to a degree, or says it's a warm day and it's warm to a degree. How will that influence the situation? That's one point.

Another point is this. I think that the best way of resolving these issues, as I've said already, is to consider particular problems, such as, if I say most students are young, and then I say most students are healthy, with the understanding that young and healthy both are fuzzy predicates, and then I ask the question "what fraction of students are healthy and young?"

I would like Prof. Lindley to come up with a probabilistic analysis of this sort of a thing, not at this point of course but at some point.

I think it was very interesting to see how it could be done. Personally, I think there would be very great complications, if it can be done at all, whereas using fuzzy logic it's a very simple sort of a thing and you get an immediate answer.

DR. LINDLEY: My response to that is if I have your simple answer I'll show you that the probability one is better.

DR. DeGROOT: Are there other comments?

DR. WISE: I address Mr. Watson's point about probability leading to complexity, and that is if you describe a problem, like Professor Zadeh did, and then you work it with probability you need to make a series of assumptions to get some answer.

If you work it with something else, you may get a very similar answer and I think the fact that you had to make the assumptions in probability to get essentially the same answer indicates that you have implicitly made them using the other theory. The mere fact that you get an answer at all indicates that you had made some assumption implicitly and you could uncover it by doing it probabilistically, assuming whatever is necessary to get the same answer.

DR. LINDLEY: I think what you said is absolutely right. I met an example recently in which an argument had been used and it turned out to be a perfectly sound argument from a probability point of view, but an assumption had been introduced at one point and you had to say to yourself, was that assumption reasonable?

The argument went through without this assumption being exposed. As soon as you did it by probability you realized an assumption, and you had to say is that assumption reasonable? In actual application it was reasonable and so the answer was all right. If I understand what you're saying, I absolutely agree with you. The probability argument exposes what you have to assume.

DR. DeGROOT: I'd like to comment on your standard example for probability. Let me say that I, like Art Dempster, agree with almost everything you say -- with everything in spirit ---- and only disagree in detail as to those here and there.

Your standard example for probability is that I evaluate or assess the probability of an event A by asking would I prefer to receive a prize contingent on A or contingent on some black ball being drawn from an urn of known composition.

I don't like that way of assessing probability. I think about the event A being nuclear war tomorrow. Would I prefer to receive $5 contingent on nuclear war tomorrow or contingent on drawing a black ball from some box, no matter how rare that black ball might be in that box, and I think I would probably go for the black box. I don't think five bucks is going to do me much good tomorrow. It's not enough to even get out of town, really.

The traditional, the classical phrase is ethical neutrality of the events A that we can assess in these terms, but many events are not ethically neutral to us, and I think there are other ways -- I still am a believer in probability, and I think there are other ways to assess the probability. I think there's a question there somewhere.

DR. LINDLEY: I don't see the ball and urn method as a sensible way of assessing probabilities any more than I think it will be sensible for us to get a van, and cart this desk out to the National Bureau of Standards and place it next to the standard yard and see how long it is.

We don't do things that way. We don't compare them with the standards. We use other techniques, and I'm sure we have to use other techniques for assessing probabilities. My point was there is a standard for probability. I don't think you would use it any more than you've used the standard for length.

DR. FORMAN: Ernest Forman from the Management Science Department at George Washington.

In making these estimates of probability, you talked about the need for a standard and you talked about different types of comparisons.

Why not make your estimates in terms of relative comparisons, so not only should you put a .4 on this and a .6 on this but say this is .6 to .4 or three to two ratio and then do whatever normalization you have to do to invoke the laws of probability.

It seems like that's a straightforward way of doing it and could also work in context of the certainty factors, instead of just putting .3 on it look at all the other alternatives and make judgments as to how likely you think these things are relative one to another.

DR. SOLAND: Professor Lindley, I wonder if you might give us a definition of uncertainty that you use, if perhaps it's useful to give a definition, and also how you interpret imprecision and whether or not you contrast it with uncertainty.

DR. LINDLEY: You're asking me about two words in the English language, is that right?

DR. SOLAND: Yes.

DR. LINDLEY: You're not asking me about two technical terms called uncertainty and imprecision?

DR. SOLAND: Yes, technical because we've talked about uncertainty and lack of precision here.

DR. LINDLEY: Precision to me is the inverse of variables of the variance in the technical sense.

DR. SOLAND: Well, but in the sense that Prof. Zadeh has quantified imprecision how would you deal with that?

DR. LINDLEY: Well, if I'm imprecise about the value of Berkeley's population, it seems to me to mean almost essentially the same as I'm uncertain about Berkeley's population. I don't see the great distinction in the English language. All I was doing in this talk was I was talking about events. I didn't go into the technical complexity of quantities. I was talking about an event. An event to me is uncertain if I do not know whether it is true or whether it is false.

For example, the event, the millionth digit of π in an infinitesimal expansion is certain. For some things I do know whether it is true or false, though its logical truth or falsity follows from what things I do know, but I don't know and I haven't done the calculation, so to me that is an uncertain event, and I will give it probability of .1.

DR. SPIEGELHALTER: You assume that all your events are going to be able to be scored in the future, that they are potentially verifiable propositions.

Do you think there is any role for propositions that aren't strictly verifiable? Let me give an example of something that may appear in expert systems for statistics. Say, that there may be a node corresponding to an assumption of normality in the data, or linearity in regression, and this for control purposes may be an important thing to establish and some idea of the compatibility of the data with that assumption is important, but it's not perhaps something that you might like to give a probability to.

I was wondering if you would say how you might deal with propositions that aren't strictly verifiable.

DR. LINDLEY: Yes, that's a very important point. When we have to do something, then of course the things that we're concerned with typically are things that can be verified but it turns out that when we do the calculations it is very convenient to bring in things whose value can never be verified.

The simplest example is the standard situation in statistics in which we have a number of random variables that are independent and identically distributed.

Now we think about the situation by bringing in things called parameters. Nobody ever knows what these parameters are -- nobody is ever going to observe them, but to bring them in is enormously simplifying in the calculation of observable probabilities.

For example, suppose I have a sequence, X_1 up to X_{10}, and I observe the values and I now want to say something about the X_{11}.

The simplest way for me to do it is to do it through parameters that I will never observe but I will observe X_{11} and this technique is really useful.

I note somebody said that my analogy with length wasn't perhaps right. Maybe it's not but here's another example. There are plenty of situations in which you cannot measure lengths.

The way you measure the distance from A to B is from A to C and C to B. You do it the long way around and there are many situations like that in probability where you can't do the thing directly. You have to invent other ways of doing it.

There are also of course questions that you can't verify. For example, did Shakespeare write Hamlet? My personal probability that Shakespeare wrote Hamlet is about .2, but nobody is ever going to be able to verify that, at least it's most unlikely, so you say does it matter?

Well, you have to turn that into, say, the following event: that it will be discovered during the next calendar year that Shakespeare did not write Hamlet.

Now that's an event that can be tested and it's an event that is of tremendous importance to the people of Stratford-on-Avon in England, because if it ever was discovered that he didn't, the tourist industry would collapse.

How would you calculate the probability that it will be discovered in the next year that Shakespeare did not write Hamlet? Well, you would have to say, supposing he didn't write Hamlet, what is the probability that it will be discovered? Supposing he did write Hamlet, what is the probability, and so on. You have have to order it conditional on Shakespeare wrote Hamlet.

In other words, you bring in an event that can never be verified in order to calculate a very important event that can be verified, namely, it is to be discovered during the next year that Shakespeare did not write Hamlet. Admittedly, the probability of that is very small, but still it's an event of supreme importance to the inhabitants of Stratford-on-Avon and can be verified.

So I think that there are events that can never be verified but are extremely useful for us to introduce into our calculations.

DR. SHAFER: I just wanted to push Dennis to address further the point about the scoring argument.

In the case of the unverifiable events, what is the relevance of the scoring argument?

DR. LINDLEY: None whatsover, but it is entirely relevant to the events that can be verified.

DR. WATSON: That was really my point as well. If I can enlarge on it, if scoring doesn't apply to unverifiable events, why should probabilities exist for unverifiable events?

DR. LINDLEY: Well, because the events that are related to them can be scored and that is enough. I'm saying this with some hesitancy. Is that right? My feeling is that it's just like distance. If I measure enough distances I can infer those are the distances. I think it's the same with probabilities. If I have enough events that are verifiable, then I can carry the argument.

DR. DEMPSTER: Would you say, Dennis, that your personal probability is .2 about Hamlet was based on evidence of some kind -- There's a kind of gulf between us here.

DR. LINDLEY: Oh yes, I have information. There are several pieces of evidence. For example, we do not have any of William Shakespeare's handwriting except one signature which strongly suggests this was a man who found great difficulty in writing.

It is very surprising that a schoolmaster should have had enough knowledge of Roman history, court behavior of the Tudors and things like that to have written the plays.

It's very amazing that this should have happened but on the other hand there is a person around who did indeed have all that knowledge and could well have written it, and that was the Earl of Oxford. There are also other people who also had the knowledge and might have written it.

So there is quite a bit of this sort of circumstantial evidence that Shakespeare was just not in the position to have written the play, he didn't know enough.

DR. DEMPSTER: These are all pieces of evidence that point one way.

DR. LINDLEY: Right. The piece of evidence that points the other way, of course, is that he put them all on.

DR. DEMPSTER: Your theory doesn't make any special distinction about the direction in which evidence points.

DR. KONG: Can I ask a question? It seems to me that in order to get that probability point two, lets say I'm a Bayesian, what I have to do is I have to look into the time before Hamlet, the piece of literature, actually appeared, and then every human being before that can write that piece of literature, so I need a prior probability distribution for every one of them. Maybe I use the uniform distribution so it's like one over so many billion; and then for each one of them I have to have like a likelihood of them writing Hamlet and then I do all the calculations and then come up with the conditional probability of Shakespeare actually writing Hamlet, the probability is .2, so we actually need a lot of numbers. We need like likelihood for all the human beings before Hamlet actually appeared, is that right?

DR. LINDLEY: No, I don't think that's right. Let me tell you a story about this. I was talking with De Finetti once, and he said to me what's the matter with you probabilists? He says you're always talking about sigma algebras. So I said, well yes, we do. We suppose these probabilities actually are in sigma algebras. Why, he said, why should I have to think about all the events in the sigma algebras? Why can't I think about some of them. Surely that's relevant here.

All I need to do is say I have an individual called Shakespeare, and I am concerned with whether that individual wrote the plays. The other possibility is that someone other than Shakespeare wrote them. I don't see why I have to consider everybody.

DR. KONG: If I do that, am I sort of hiding some kind of assumptions because every human being apart from Shakespeare, at least it's possible that they did write them.

DR. LINDLEY: But if you now ask me the question -- you are going to find some specific person who was alive, say just about the same time as Shakespeare, and ask me what is the probability that he wrote Hamlet, then indeed I shall have to do the sort of calculations you suggest. But whilst my question is did Shakespeare write it or did he not, I don't see why I have to engage in this.

DR. KONG: It all depends on other possibilities, because like if Shakespeare is the only human being that existed, of course the probability has to be one because nobody else can write it. Of course it's very unlikely for him to do it but if he is the only one of course he has to be the one.

DR. LINDLEY: All the possibilities have been eliminated.

DR. KONG: Right, so it seems to me that in order to get the probability you still have to consider the alternative. Of course we don't consider each individual by itself but we are sort of making some kind of assumptions there. Sort of coursening all those individuals together.

DR. DeGROOT: What you're stating is the nuisance parameter problem and it is a fundamental problem in Bayesian statistics and Bayesian methodology in general.

What I sense, Dennis, in answering is that in some circumstances one can mentally if no other way simply integrate out all that's not necessary to assess distributions on an entire high dimensional space, if you're only going to be interested in one particular event and the nature of the information that you have makes it possible to directly assess the probability of that event.

DR. KONG: The thing is, theoretically I do have a prior and theoretically it may be subjective but I do have some kind of likelihood I can construct and I can look at each human being and then theoretically I can use the Bayesian formula and get the answer.

DR. DeGROOT: All that is true. I agree with that. In fact Dennis challenges you can do that. It seems to me the way he's bringing forth his evidence is almost perfect for belief functions. I have no idea at all how to do it Bayesianly so Dennis has to tell us.

DR. KONG: It seems to me that if it's like belief function its like .2. But if its Bayesian it seems I have a prior I have to like, I must have constructed the probability of .2 out of these two pieces of evidence. I don't think anyone can state clear out what is the prior and what is the likelihood of each individual human being. It is sort of impossible. It has too many parameters.

DR. SHAFER: Not even too many. Tversky has done some work where he's gone through and shown that people do not maintain this coherence that was mandated by Bayesian. You give them simple examples and ask them for priors and they won't agree with Bayesian statistics.

How do you capture these things? How do you get these numbers? People don't think in these terms.

DR. DeGROOT: I'll let you handle the hard ones, Gentlemen. I'm just the moderator here.

DR. LINDLEY: My only difficulty in answering that question is that you wouldn't let me talk for an hour and a half on it.

The first thing is that it seems to me preposterous to expect people to do these things when they haven't had any training in probability.

To expect people to be able to do this sort of thing without training in probability does seem to me to be ready rather absurd, so my attitude to all these Tversky results is to say, well, yes, I'm not surprised, how could they succeed otherwise.

The other point is of course this is not a descriptive theory. It is a normative theory. It's how you would wish to behave if only you could do it.

Well, it may be, it may be, that after we've worked with this for 20 years we discover that you just can't do it. Maybe so, and it would be a great pity, but I think we ought to try.

Just imagine that we weren't in 1984, we were in 1684 and that Isaac Newton's theory was hot off the press, Newtonian mechanics.

It would be absurd to have turned around to Newton and said, "Oh, your theory is absolutely useless because we can't measure masses and accelerations and these things accurately enough to do the job."

What you did was to develop ways of measuring those things and I think that the logic here is so strong that what we now need to do is to put an investment into whether these things can be measured.

To expect a human being to be able to do it without any training at all seems to me to be rather unsatisfactory.

DR. LINDLEY: I am delighted by the fact that people can't do this. If they could do all this thing naturally, then I would be out of a job.

(Laughter.)

It is because they cannot do these things that we probabilists have a potentially marvelous tool. People cannot do these things so now we have something to help them.

I think that is great. I think this is one of the greatest things of the 20th century. If they could do it naturally then I wouldn't be here.

PROBABILISTIC EXPERT SYSTEMS IN MEDICINE: PRACTICAL ISSUES IN HANDLING UNCERTAINTY

David J. Spiegelhalter
MRC Biostatistics Unit
Medical Research Council Centre
Cambridge, England

PROBABILISTIC EXPERT SYSTEMS IN MEDICINE: PRACTICAL ISSUES IN HANDLING UNCERTAINTY

1. INTRODUCTION

The first problem in discussing "uncertainty in expert systems" comes in defining our terms. We shall take a fairly narrow view of 'expert systems' and consider only programs that contain a 'knowledge-base' of interrelated propositions, represented in such a way as to be usable by an 'inference engine' to make some type of human-like judgment concerning 'data' from a new individual. Generally, such a system is hoped to be able to justify its judgments in a manner comprehensible to the user, to allow updating of its knowledge base in response to experience, and to be based largely on the expressed opinion or observed practice of one or more 'experts.' We shall concentrate specifically on 'diagnostic' systems whose aim is to weigh evidence concerning possible outcomes whose status is currently unknown - this includes systems for prediction as well as the more standard classification programs.

Within this context the term 'uncertainty' is commonly used in a broad spectrum of qualitative ways: for example, to describe incompleteness in the knowledge base ("I don't know what we should think if X occurs"), to describe doubt about the structure of the knowledge base ("I'm not sure whether it is reasonable to assume X and Y are independent") to qualify logical implication ("X --> Y with certainty P"), to describe imprecision about the qualification ("I'm not sure what P should be"), to describe ignorance concerning the current individual ("I've no idea whether X is true or not"), and even, occasionally, to describe the probability that a proposition is true. 'Uncertainty' has also been used in describing the extent to which a proposition is true, although this is an area which appears to fall within fuzzy reasoning. There is often an interpretation in terms of degree of support for a hypothesis, expressing the matching or compatibility of the observations with those expected were a hypothesis true. Each of these ideas has been discussed from different professional perspectives, and the view that any deviation from a self-contained logical system is 'uncertainty' has led to much general discussion of multi-valued logics of which probability is only one example. In this paper it is argued that probability theory does indeed have a strictly limited role, but that within these limits it can adopt many of the desirable characteristics of methods adopted by others.

We shall concentrate on practical, rather than philosophical, issues concerning the way uncertainty is handled in existing programs, and do not consider in detail either the representation of knowledge or the control of the program. Published examples motivate the search for a methodology that satisfies a number of demands, and three current projects will then be used to illustrate some specific aspects of the attempt to use probabilistic methods in as effective a way as possible. Finally, an attempt is made to bring the argument together into a prospect for future developments.

2. DEMANDS MADE OF A CALCULUS

The particular complexity of many medical problems has challenged the notion of a rigorous unified treatment of uncertainty and, in general, ad hoc quantifications have been used to measure evidence for various possible underlying hypotheses (Szolovits and Pauker, 1978). The complex interrelationships between disease processes and manifestations has led to various systems for propagating degrees of certainty and combining evidence from different sources - PIP (Pauker et al, 1976) and INTERNIST/CADUCEUS (Miller et al. 1982) both essentially score hypotheses using evidence from current symptoms that support a hypothesis, which is discounted by a score expressing absent symptoms that would be expected, and a score expressing present symptoms that would not be expected. MYCIN/EMYCIN use a more modular structure in which 'certainty factors' are propagated, while CASNET/EXPERT (Kulikowski and Weiss, 1982) propagates 'weights' through a causal network. A statistical system such as that of de Dombal et al, (1972) begins with 'knowledge' derived from a data base, but the simplistic independence assumptions made in combining evidence (although effective in discrimination) ensure that the 'certainty' propagated is not expected to be interpretable as a probability - the same holds for the 'Bayesian' updating technique in PROSPECTOR (Duda et al, 1976). Fuzzy reasoning (Adlassnig, 1980, Fieschi et al, 1983) has also been used as a means of capturing the ill-defined nature of many clinical terms.

We can identify a number of considerations that have led to the procedures that have been adopted and that are currently being researched. Strongest has been the claim that a single probability of a hypothesis, even if it were based on extensive data, is not sufficient to convince a clinician: the evidence on which to base a conclusion must be retrievable, to enable conflicts and doubtful contributions to be identified. A particular case of this demand for justification is the situation where little relevant data is available and there is essentially ignorance concerning the possibility of a hypothesis. This arises particularly often in medicine due to the hierarchical, taxonomic structure of disease descriptions in which evidence may be available which supports a general disease category but gives no indication of the relative plausibility of the sub-categories of disease. Thus the hierarchical hypothesis structure is viewed as a natural justification for ranges of uncertainty, for which a number of schemes exist (see, for example, Quinlan (1982)). The demand that individual contributions of pieces of evidence should be identified, and that evidence should be able to focus on groups of diseases without distinguishing within that group, has led naturally to the study of the possible role of belief functions in medicine (Gordon & Shortliffe, 1984). Much attention is now being paid to solving the accompanying computational problems and making some allowance for dependencies between sources of evidence. The concept of 'discounting' in belief functions could also be seen as a means of allowing for doubt about the precise numbers to be placed on evidential statements.

To summarize: current interest is focused on schemes that can propagate measures of uncertainty through complex relationships often defined on a hierarchical structure, that can identify conflicting evidence and lack of evidence, and can cope with incoming data that do not follow a pre-defined order. The reasoning process should be justifiable and fairly intuitive, and allowance for imprecise specification of numerical relationships would be an advantage.

While the above desiderata appear admirable, we feel there is an important item that has been largely ignored in practice. This concerns the operational meaning of the quantities which express uncertainty. In the following examples we describe attempts to retain 'meaning' while responding to demands and constraints made by the real practical problems of interest.

3. EXAMPLES OF PROBABILISTIC ANALYSIS

GLADYS - The GLAsgow DYSpepsia system

GLADYS is a program designed to interview patients presenting to a clinic with dyspepsia, and provide a reasoned probabilistic diagnosis based on the symptoms alone. It was developed at the Diagnostic Methodology Research Unit at Glasgow, and runs on a microcomputer with a special keyboard to record patient responses. The control of the interview is strictly algorithmic, in that branches to more detailed interrogation are taken depending on the results to 'trigger questions,' and the interview has been found to be accurate and acceptable (Lucas et al, 1976). The responses are analyzed according to a scoring system derived from a modified logistic regression technique, described in detail in Spiegelhalter and Knill-Jones (1984), of which certain aspects are relevant to the issues raised in the previous section.

Firstly, there is a real need to deal with hierarchical disease structures, in which for example, certain features may discriminate the generic class 'peptic ulcer' (PU) from other diseases, while other items of information are relevant to discriminating duodenal from gastric ulcer (GU) within the peptic ulcer class. This is accomplished by calculating probabilities conditional on the branch in the hierarchy and then multiplying downwards to obtain the overall probability: for example, we calculate $p(GU|PU)$ and $p(PU)$ from which $p(GU) = p(GU|PU)p(PU)$.

Secondly, the scoring system allows explanation of the final probability in terms of the contributing pieces of evidence. For example, a patient described in Spiegelhalter and Knill-Jones (1984) provided the following evidence relevant to a diagnosis of gallstones:

Evidence FOR gallstones		Evidence AGAINST gallstones	
History less than 6 months	77	Pain not severe enough to warrant emergency call to doctor	-43
Pain comes in 'attacks'	177	Pain does not radiate	-38
Can enumerate attacks	63		
Attacks produce restlessness	31		
Pain in right hypochondrium	77		
	425		-81

Balance of evidence +344 (Total evidence = 425 + 81 = 506; conflict ratio = $\frac{506}{344}$ = 1.5)

Initial score -300 (Corresponding to prevalence of 4.7%)

Final score 44 ≐ 61% chance of gallstones

Some explanation of the above 'explanation' is necessary. The scores given to findings are 100 log (likelihood ratios) adjusted, roughly speaking, for correlations between items of information. Thus the initial score of S = -300 is transformed to a prior probability p = 1/{1 + exp(-S/100)} = .047, which is simply the inverse of S = 100 log {p/(1 - p)}. The 'conflict ratio' is a rough measure of how much the total evidence obtained contradicts itself : a high ratio, say above around 2.5, suggests the clinician should check some of the important questions. The initial score is based on a prevalence in an urban clinic and could be altered depending on circumstances. The scores come from analysis of a data base of 1200 cases and the statistical modeling means the final probabilities are reasonably 'well-calibrated', in that of patients presenting as above, around 60% should turn out to have gallstones as a major cause of their symptoms. There is, however, no reason why the scores should not be subjectively assessed provided one could ensure the predictions had similar properties of calibration.

Thirdly, imprecision of the quantification could be incorporated by placing standard errors on the predictions; the above example has a standard error of 42 on the final score corresponding to a rough 95% interval of (.40, .78) on the predictive probability. **Finally,** ignorance may be viewed 'retrospectively' in terms of **the 'total** evidence' received either for or against a proposition. However, as suggested in Spiegelhalter and Knill-Jones (1984), we may also quantify 'prospective ignorance' in terms of the results that may occur when the data of which we are currently ignorant becomes available. This concept translates into calculating the predictive distribution of the possible final probabilities that may be ascribed to a disease. For example, before an interview starts, Figure 1 displays the distribution of final scores for gallstones among those with and without gallstones. Tukey

(1984) recommended that such distributions should be included as part of the explanation facilities.

Thus before an interview, a patient has a fairly precise probability of gallstones (95% interval .03, .07) but one based on an ignorance reflected in the wide distribution of feasible probabilities that could be taken on when data become available; while at the end of the interview, there is a relatively imprecise probability with a 95% interval of (.40, .78), but no remaining ignorance within the bounded context of the system.

We would not normally consider GLADYS as an 'expert system' since it does not use knowledge representation techniques derived from AI, it is not based on 'expert opinion' and it does not operate interactively. However, many of our aims match those of 'classic' expert systems, except that we are determined to remain, as far as possible, within a probabilistic framework.

A diagnostic system for chest diseases

A group at the Chest Clinic, Westminster Hospital are developing a system for probabilistic diagnosis of patients presenting with a normal chest X-ray. The system uses simple independent Bayes updating assuming mutually exclusive disease categories, and our only concern here is with the subjective probability assessments on which the system is initially based. The consultant physician has been required to assess prior probabilities for each of the diseases conditional on the age group of the patient and the main presenting symptoms, as well as the probabilities of the secondary symptoms conditional on each of the diseases. Around each probability he was required to place an interval reflecting his confidence in the point probability. By viewing this range as an approximate 90% interval around a binomial probability one can derive a rough 'implicit sample size' on which his judgment of each probability has been based. These measures of imprecision are currently not propagated through the consultation, although Rauch (1984) suggests ad hoc methods of doing this while allowing for correlated judgments. However, the implicit sample sizes allow the probabilities to be stored as a fraction r/n, and where a confirmed case with the relevant symptom is found the probability may be updated to $(r + 1)/(n + 1)$. This emphasizes that probabilistic systems may be based on subjective opinion, and yet a rational means of allowing that opinion to learn from experience is easily available.

IMMEDIATE - a system for general practice

In contrast to GLADYS, IMMEDIATE is a rule-based AI system written in PROLOG which is being developed by a group centered at the Medical Computation Unit at the University of Manchester. It is designed to support certain activities of general practitioners and its control philosophy is described elsewhere (Dodson & Rector 1984).

Two aspects of its development are of interest here. Firstly, although the knowledge structure and uncertainty propagation bears some resemblance to that of PROSPECTOR, a deliberate aim is that the probabilities should be made to 'cohere': thus initial probability judgments should form a valid joint distribution, and, as data arrives, uncertainty be propagated in a way that retains its interpretation as subjective probability. Secondly, part of the control mechanism is based on a range of 'ignorance' or 'evidence availability' that is an explicit calculation of the maximum and minimum probabilities of a proportion that could be achieved when further information becomes available. This may be seen as a summary measure of the predictive distributions of final probabilities described under GLADYS. Explicitly calculating the range of potential probabilities of a proposition helps towards an assessment of the importance of establishing relevant patient characteristics, which in turn ensures that the clinician is informed as to the most telling questions to ask.

4. DISCUSSION

The preceding section is an inadequate indication of the work currently being carried out in probabilistic systems, and we have only been able to mention aspects according to their capacity to illustrate the practical implementation of important issues in the handling of uncertainty. In this section, we attempt to summarize these issues, with the aid of examples drawn from the systems introduced above.

Status of Propositions

It is clearly preferable that all propositions in a system are crisply defined and, at least theoretically, verifiable at some point in the future, as required by Smith (1965) or de Finetti (1974). Nevertheless, the inevitable imprecision of statements (e.g. "the pain is relieved by food") makes it tempting to allow degrees of truth of propositions and adapt a fuzzy calculus. It should, however, be emphasized that it is not the true state of the world to which the system has access, but the assertion of the state of the world (The patient has replied YES to the question "Is the pain relieved by food?"), and this is necessarily made 'crisp' by the restricted means one has to put information into the system (e.g. just a YES/NO button). An expert system can therefore force the user to be categorical in his assertions, although we acknowledge that user demand for qualifications of 'degree' may create the need for an alternate calculus to deal with 'partly-true' propositions.

A statistician may tend to view a knowledge-base as a set of related 'nodes', each corresponding to a random variable which may take on a number of mutually exclusive and exhaustive values. The 'rules' attempt to define a distribution on the variables. For control purposes, however, it may be necessary to have 'action' nodes which correspond to conclusions on which further analysis is conditioned. These may well not be strictly verifiable propositions; for example, in a system designed for statistical analysis, there may be assertions of 'normal errors' or 'linear relationship'. Strictly speaking a decision-theoretic argument should be used for any interim decision made

in the control of a consultation, but this is not usually practicable. Instead, it may well be reasonable to adopt a calculus of 'compatibility' or 'degree of support' for a hypothesis for which a probability is not well-defined.

Knowledge Representation and Explanation

We feel that probabilistic methods can handle hierarchical taxonomic structures without extending into belief function methodology. There is, however, a great need for further work in coherent assessment and propagation of probabilities through the network structures arising from rule-based systems. The graphical representations of certain log-linear models described by, for example, Wermuth & Lauritzen (1983) appear to be relevant, with propagation schemes extended from those of Kim & Pearl (1983). Subjective judgments may be deliberately over-specified to allow for identification of incoherence due to poor assessments or weak modeling, or underspecified and 'padded out' using, for example, the maximum entropy methods of Cheeseman (1983) and Konolige (1982). Using such a structure and explanation facilities similar to GLADYS, one should be able to fulfill the aim, described by Dempster (1985) of justifying quantified judgment explicitly in terms of the sources of evidence.

Intervals and Probabilities

As we emphasized in discussing GLADYS, two types of range of probability must be distinguished. The first, due to inadequacies in the knowledge base, concerns the imprecision in the quantifications. This may be represented by a standard error or even a fuzzy qualifier in the manner suggested by Freeling (1981), but in either case the range represents a type of automatic sensitivity analysis conditional on the data already obtained. This interval tends to widen as more data come in.

This should be contrasted with an interval based on ignorance concerning the current case, and one way in which this can be defined is in terms of the probabilities that could be taken on when the unknown data, denoted X, becomes available. If D represents a disease with current probability p(D), then the predictive distribution of the eventual probability $p(D|X)$ may either be fully calculated as in GLADYS or summarized by its range as in IMMEDIATE. We note that

$$
\begin{aligned}
E[p(D|X)] &= \int p(d|X)p(X)dX \\
&= \int p(X|D)p(D)dX \qquad \text{by Bayes theorem} \\
&= p(D)
\end{aligned}
$$

Hence our current probability may simply be thought of as the mean of the distribution of possible final probabilities. This distribution narrows as the consultation proceeds.

In this way ignorance is explicitly defined in terms of the X that we do not yet know. In real life, X is unbounded and so such a calculation is unreasonable, but it is important to note that an expert system is bounded and so can always explicitly state what information is missing, provided a suitably efficient search routine is available.

Operational Meaning

Our practical experience has strongly influenced us towards establishing operational meaning to our expression of uncertainty. This has three stages: firstly, the inputs, based on either real or 'imaginary' past data, must have sufficient interpretation to allow informed argument. Clinicians often disagree strongly about frequencies, but we have found the resulting discussions illuminating and constructive: the problems of agreeing on numbers with no verifiable interpretation is vividly illustrated in the fascinating transcript of an argument concerning 'certainty factors' contained in the recent book on the MYCIN projects (Buchanan & Shortliffe, 1984). Secondly, preserving operational meaning in the propagation of uncertainty requires attention to the coherence of the assessments when placed in a large, complex knowledge-base. Finally, the outputs need to have an externally verifiable interpretation in terms of their 'calibration' against experience. Such calibration is not part of the axioms of subjective probability, but we have found an enthusiastic response from clinical colleagues when they find the predictions provide reasonable 'betting odds'. Of course, a system may process information solely with the aim of providing a, possibly ranked, set of alternatives with some attached measure of evidential support. However, if a system is to be used to guide the choice of an option, this seems to be inadequate. In fact, a subjectivist statistician may view a diagnostic expert system as a 'coherence machine', which takes in relevant information, and throws out acceptable betting odds on future events.

Finally, perhaps the most important reason for interpretable quantification is the need for learning. As we have illustrated with the chest disease system, updating of subjective probabilities is feasible and should provide a convergence of opinion that may overcome local biases which may otherwise render a system unacceptable.

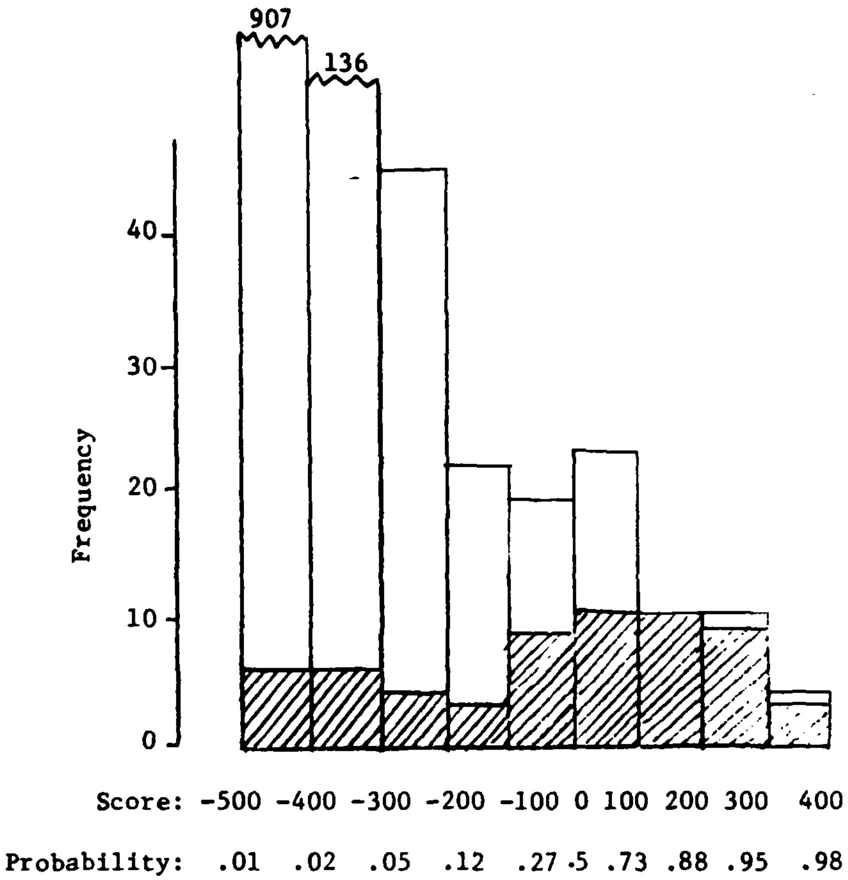


Figure 1. Empirical predictive distribution of final score on gallstones:

1119 cases of 'not gallstones'

57 cases of gallstones (shaded)

BIBLIOGRAPHY

Adlassnig, K.P. (1980). A fuzzy logical model of computer-assisted medical diagnosis. Methods Inf. Med., 9, 141-148.

Buchanan, B.G. and Shortliffe, E.H. (1984). Rule-based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project. Reading: Addison-Wesley.

Cheeseman, P. (1981). A method of computing generalized Bayesian probability values for expert systems. In proceedings of 8th International Joint Conference on Artificial Intelligence. Karlsruhe, West Germany, p. 198-202

de Dombal, F.T., Leaper, D.J., Staniland, J.R., McCann A.P. and Horrocks, J.C. (1972). Computer-aided diagnosis of acute abdominal pain. Brit. Med. J., 2, 9-13.

de Finetti, B. (1974). Theory and Probability, Vol. 1, London, Wiley.

Dempster, A.P. (1985). Probability, evidence and judgment. In Bayesian Statistics 2 (J. Bernardo et al, eds.) (to appear).

Dodson, D.C. and Rector, A.L. (1985). Importance - driven distributed control of diagnostic inference. In Research and Development in Expert Systems (Bramer, M.A. ed). Cambridge University Press: Cambridge, England..

Duda, R.O., Hart, P.E. and Nilsson, N.J. (1976). Subjective Bayesian methods for rule-based inference systems. Proc. AFIPS. Nat. Compt. Conf.,47,1075-82.

Fieschi, M., Joubert, M., Fieschi, D., Botti, G. and Roux, M. (1983). A program for expert diagnosis and therapeutic decision. Medical Informatics, 8, 127-135.

Freeling, A.N.S. (1981). Alternative theories of belief and the implications for incoherence, reconciliation and sensitivity analysis. Decision Science Consortium.

Gordon, J. and Shortliffe, E.H. (1984). The Dempster-Shafer theory of evidence. In Buchanan & Shortliffe (1984), 272-292.

Kim, J.H. and Pearl, J. (1983). A computational model for causal and diagnostic reasoning in inference systems. In Proceedings of 8th International Joint Conference on Artificial Intelligence Karlsruhe, West Germany, p. 190-193.

Konolige, K. (1982). Bayesian methods for updating probabilities. Final Report, Project 6415, SRI International.

Kulikowski, C.A. and Weiss, A.M. (1982). Representation of expert knowledge for consultation: the CASNET and EXPERT projects. In Artificial Intelligence in Medicine (Szolovits, P. ed) Colorado: Westview Press 21-55.

Lucas, R.W., Card, W.I., Knill-Jones, R.P., Watkinson G. and Crean, G.P. (1976). Computer interrogation of patients. Brit. Med. J. 2, 623-625.

Pauker, S.G., Gorry, G.A., Kassirer, J.P. and Schwartz, W.B. (1976). Towards the simulation of clinical cognition: taking a present illness by computer. Amer. J. Med., 60, 981-986.

Quinlan, J.R. (1983). Inferno: a cautious approach to uncertain inference. The Computer Journal 26, 255-269.

Rauch, H.E. (1984). Probability concepts for an expert system used for data fusion. AI Magazine, 55-60.

Smith, C.A.B. (1961). Consistency in statistical inference and decision (with discussion). J. Roy. Stat. Soc., B, 23, 1-25.

Spiegelhalter, D.J. and Knill-Jones, R.P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion). J. Roy. Stat. Soc., B, 147, 35-77.

Szolovits, P. and Pauker, S.G. (1978). Categorical and probabilistic reasoning in medical diagnosis. Artificial Intelligence, 11, 115-144.

Tukey, J.W. (1984). Discussion of Spiegelhalter and Knill-Jones (1984).

Wermuth, N. and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. Biometrika, 70, 537-52.

TRANSCRIPT OF ORAL PRESENTATION BY DAVID SPIEGELHALTER: PROBABILISTIC EXPERT SYSTEMS IN MEDICINE, PRACTICAL ISSUES IN HANDLING UNCERTAINTY

DR. SPEIGELHALTER: I am an applied medical statistician working in the MRC Biostatistics Unit in Cambridge. But I have a long interest in decision making, since I was reared and indoctrinated at University College, London under Dennis Lindley and Adrian Smith. Recently I have become involved in a number of projects where people have been attempting to apply the techniques, or at least some of the ideas, of AI in medical diagnostic problems.

The first problem in talking about uncertainty in expert systems is defining expert systems and defining uncertainty. And we have had quite a bit of talk about uncertainty yesterday, but not a lot really about what expert systems are.

Firstly I want to say what I feel are the aims of expert systems in medicine. My examples will all be taken from medicine although I hope that a lot of the things I talk about are applicable much more generally. In particular, I am interested in "Classification" types of diagnostic expert systems, not critiquing systems designed to comment on a proposed course of action. They are generally there in order to make some sort of judgments about some unknown aspect of some individual person, which might be a diagnosis or a prognosis.

The basic structure consists of a knowledge base, kept separate from an inference engine, which controls the process by which the knowledge is used in order to make some sort of judgment on a new individual from whom data is obtained. These are just sort of buzz words which can mean all sorts of things in different applications.

What is often considered a necessary characteristic of an expert system is that they should be able to justify their reasoning by making the process by which they obtain their judgments explicit, interpretable, and understandable to the user. They must be able to justify their conclusions.

To some degree the knowledge base will be based on expert opinion. Whether that is quantitative expert opinion, or only qualitative expert opinion in relation to the structure of the knowledge base, will depend, again, on application to application.

People in AI say that you should be able to update a knowledge base from experience, especially from its failures as well as from its successes, and we should be able to learn in some way. And the control of the consultation (consultation perhaps should be in quotes) will be largely based on some heuristic techniques which attempt to bear some resemblance to how an expert may attempt to solve the problem. So these are some very general phrases that say what might be the basic aims that

we are trying to fulfill.

Now, I am going to talk completely about applications, about past applications and about work in which I am involved. These applications do cast light on some of the theoretical issues that were being discussed yesterday.

Before getting to uncertainty, I would like to talk about knowledge representation. Here is one example of representation of medical knowledge with an expert system, an old one, the CASNET system developed at Rutgers. In this case an explicit attempt is made to represent the physiological causal knowledge that clinicians have about the disease glaucoma. (see slide 1)

In particular, we can see that there is a plane of observations. These are the actual variables, the data that can be observed from the individual. And these are considered as being caused by some unobservable patho-physiological states which in turn are caused by the deepest level of the underlying unobservable disease states. It is these that we are really interested in, but we can only observe the diseases through this intermediate layer.

There are all sorts of ways statisticians might see that as being in terms of latent class models, in order to allow for dependence by putting in intermediate states. A less structured system is something like MYCIN, where the knowledge consists of little chunks of production rules which relate to particular groups of findings. There is some underlying structure but not very much.

Getting onto some things that are a bit more complicated, this is just a part of the structure of INTERNIST or CADUCEUS. This may just look like a lot of jibberish. But I find when I put this up in front of clinicians, after a bit they start seeing that this does make sense. There are certain aspects of this that relate to what Glenn Shafer was saying yesterday. (see slide 2)

First of all, CADUCEUS is this massive system with 500 diseases and 3000 possible symptoms and is supposed to cover most problems in internal medicine. But looking at the knowledge structure that is used, I think, is important, because this relates to many problems in medicine. The diseases, which are the blocks here, the pathological and nosological descriptions, can be related very often in some hierarchical taxonomic structure.

So we have here this connector which is a subclassification. We have hepato-cellular involvement, of which a particular type is fibrotic hepatocellular involvement, of which a particular type is cirrhosis of which a particular type is biliary cirrhosis. So very often we do get this hierarchical representation of the underlying diseases.

Superimposed on that you have a causal network where in particular, there is a "caused by" link that is obtained so you can see the upper gastrointestinal hemorrhage can be caused by the portal hypertension. So what that is doing, having an underlying taxonomic disease structure with a causal network superimposed which affects different levels of that structure. That is exactly the picture that Glenn Shafer put up yesterday where particular items of evidence may tell you about particular levels of the disease, and that recurs many times.

So that is an example of the sort of flexibility, the structuring that people try to put into expert systems in medicine.

What about uncertainty? Well, before becoming technical about it, I will start out in a linguistic way describing how people might use the word "uncertainty" in describing attempts to represent expert knowledge. Here are just a few of them and you can just keep on going on the list. One type of uncertainty has to do with incompleteness of our knowledge, that in a particular set of circumstances we have not provided for the inference. What do we think if X occurs?

Another type of uncertainty could be doubt about actual qualitative structure of that knowledge base. Is it reasonable to assume that two things are independent or not? Another type of uncertainty, and generally the type that is very often discussed in relation to production rules, is sometimes called the degree of implication. That is, X implies Y with some attached measure or uncertainty factor, or whatever.

Another type of uncertainty is the imprecision about this quantification. Another type of uncertainty is ignorance, where you have no information about whether a particular disease is present or not. This is often used in hierarchical structures, because you may have information at one level of hierarchy and not be able to say anything lower down the hierarchy - this example was shown yesterday.

Of course, one might even talk of the probability in terms of, say, betting odds on X being true. There are other ways in which the word uncertainty is used. It is often used in terms of the imprecision of a proposition, the degree of truth, or the extent to which X is true. It is often used in terms of degree of support of the evidence for an underlying disease. I am not saying the probability of the disease, but just some degree of matching or some degree of compatibility. It is often used as well about uncertainty in terms of action, e.g. "I don't know whether it is reasonable to assume X from now on". This is in terms of control strategies. There may be an uncertainty about going down a particular path i.e., "Is this a reasonable thing to do or not?" Cohen's endorsement work seems to use this interpretation a lot.

I just want to show the ways in which these terms are used. I have found in going to meetings that it is sometimes very difficult to work out what people are talking about because of the wide range of descriptions people use. Different people from different professions have different ways of approaching the subject.

People from the natural language area may try to imitate how people use these phrases. So what I would like to emphasize here is that I am using uncertainty in a very particular way. I am talking about probability, but argue that probability is more flexible than has usually been considered.

Let's look at uncertainty and how it is handled in some working systems. Here is a trivial system but one that is actually used in a number of British hospitals. A patient comes in with acute abdominal pain and the casualty officer, in the accident and emergency department, rings the relevant symptoms on the form. The rings go through on a sheet of paper with numbers on it. He types the numbers into a micro - they have been using an Apple or Commodores. At the bottom of the printout comes the probability that they have appendicitis, that they have pancreatitis, et cetera. This is a statistical system, which generates probabilities and is in use. There is a big trial going on that has just finished, where 16,000 people have gone through this system to see whether or not it makes any difference at all to their health. We are still trying to work out whether it does.

I was not responsible for the design of the system, I am only involved in the evaluation. But the way it works is usually known as Idiot's Bayes in the trade, or conditional independence. It is using Bayes' theorem, assuming conditional independence within disease groups. Essentially, the knowledge in the system, which is barely worth the name, it is just a matrix of conditional probabilities, saying for each disease what proportion of them have any particular characteristic that will be shown, based on past data. (see slide 3)

DR. COHEN: Is this assuming the diseases are mutually exclusive?

DR. SPIEGELHALTER: Yes, this is assuming the diseases are mutually exclusive and exhaustive and symptoms are conditionally independent given the diseases. About 1000 papers have been published using this Idiot's Bayes method, but there has been very little effect on clinical practice. So someone with appendicitis, their prior probability is .26 and these likelihoods are put in, for example 23 percent of appendicitis patients have right lower quadrant pain, in fact. This is just three symptoms, but in fact there could be more going in here. They are all multiplied up. You get a total which is normalized down to one and this comes out as the probability. So that is the simplest model that is very frequently used in ---

DR. SINGPURWALLA: What is the data on the new individual?

DR. SPIEGELHALTER: These are the findings on the new individual. If someone comes in and they are female and they are age 16, they have got right lower quadrant pain, et cetera, and there are lots more findings than this.

DR. SINGPURWALLA: Are those numbers, .49 and so on, the proportion of females?

DR. SPIEGELHALTER: No, the proportion of people with appendicitis who are female. So the model says you multiply together the likelihoods and the prior, normalizing to one and that is the approach. It is very crude, over-simplistic statistical modeling but one that is very often done.

And the explanation will give you the symptoms that you typed in, and the probabilities attached to these diseases. Now, one might see that the symptoms are not particularly independent, given the disease. For example, the questions "did they have previous abdominal surgery" and "have they got an abdominal scar" are both there and you sort of think they might be slightly related, even within disease classes. So the effect is that you might stick probabilities in and you may process them using some sort of statistical model, but by the time they come out, the numbers don't resemble probabilities in any sense that you would want to bet on them, and in general because of double counting of evidence the probabilities are too extreme. You have got a lot of 99 percent chances of appendicitis, but less than 99% are correct.

So you can't really trust these numbers. They provide a ranking, some measure of evidence for the disease, but they are not really probabilities by the time you are done.

The knowledge base of that acute abdominal pain system, is simply a disease and a lot of conditionally independent symptoms with no additional structure. PROSPECTOR will provide a deeper structure in which you have a series of implications between nodes, drawn up as an inference network, but it essentially tries to use a similar calculus where, notice in that previous system the actual updating mechanism was multiplying by likelihood ratios. In PROSPECTOR, similarly the numbers attach to each link between one finding f and a conclusion c is a likelihood ratio, so if f is true you multiply the odds on c by two and if f is false you multiply the odds on c by another factor. If the finding has a probability of being true you take some sort of weighted average.

So it is a sort of vaguely Bayesian system, but also in this case by the time you have gone through the network the numbers don't really resemble anything you would want to call probabilities.

It was mentioned yesterday, the MYCIN people were very interested in looking at belief functions and here is an example that they have talked about. They would like to use belief functions within MYCIN, as a response to the hierarchal nature of the disease hypotheses and the fact that certain pieces of evidence hit different levels of the hierarchy.

So here you have got one particular group of diseases. These break down in a hierarchy, but you might have one piece of evidence that gives a weight just to this pair of diseases, say intrahepatic jaundice and another piece of evidence that suggests extra-hepatic. The rest of it we leave unassigned. This will give a probability mass over this hierarchy and that leads to a range of belief, belief in the possibility of any particular hypothesis which is just the sum of the masses on elements below it and the upper point is one minus the sum over the elements that don't contain the subset of interest. (see slide 4)

DR. SINGPURWALLA: I am sorry for this point of clarification, I don't know what all of this means. Which is the disease, the thing at the top? Which is the figure you are interested in, the bottom or the top?

DR. SPIEGELHALTER: These are the separate elements of the disease. You can assume it is one of these. You don't know which, but they decompose naturally into a hierarchy of subsets.

DR. SINGPURWALLA: So the patient comes in complaining about gallstones.

DR. SPIEGELHALTER: No, the patient comes in complaining and you have narrowed it down to chotestatic jaundice. So you assume it is within here somewhere.

DR. SINGPURWALLA: That is made up of four parts?

DR. SPIEGELHALTER: Yes, four possible diseases that are labeled chotestatic jaundice, four possible components.

DR. SINGPURWALLA: But I am thinking the patient would come in complaining about gallstones or something like that.

DR. SPIEGELHALTER: No, these are the unobservable diseases with a causal network where pieces of evidence affect different levels of this hierarchy.

DR. SINGPURWALLA: And your goal is to catch the top?

DR. SPIEGELHALTER: No, your goal really will be to identify one of these particular diseases at the bottom, but your evidence may only in fact tell you that there might be evidence that supports the hypothesis that it is either hepatitis or cirrhosis, but there is no evidence there to tell you what the individual disease might be, and this is a very common structure.

So what have people been trying to do with uncertainty in expert systems? What are the practical objectives for the calculus of uncertainty? There are various things that come out of all of this discussion. First, a single number applied to a hypothesis is generally considered insufficient and I would agree with that. Just a system that goes crunch, crunch and bangs out "probability is .73' is not considered

acceptable and I don't blame anybody for not considering that acceptable.

So what do people want to be able to do? There are a lot of claims that they want to be able to cite the sources of contributing evidence. In MYCIN, there may be a trace of the rules of being fired, showing how that conclusion was reached. In particular in INTERNIST, you can identify what evidence supports the hypothesis and what evidence is against the hypothesis so one can see where there is conflict.

They want to be able to cope with hierarchical hypothesis structures. They would like to propagate through networks when you have got data coming in in a very sporadic fashion, bits of information coming in all over the place. They might like to make imprecise specification of the quantities in the system because people aren't too happy about giving single numbers.

There is also an idea that the point value for an uncertainty may be considered unacceptable, and a number of reasons why one might prefer a range or some sort of curve over that value have been suggested.

The first is what I would like to call due to "ignorance." This is related to what Morris DeGroot said yesterday about the .2 for Shakespeare writing Hamlet where, one would like to know the sensitivity of that .2 to the coming in of new information. One essentially says if that .2 is just off the top of your head, based on considerable ignorance, then one might feel that somehow that .2 is only the center of a number of possible probabilities could be attached to that proposition. I would like to get into that idea a lot more later.

This range is due to limited evidence on a new case, once in general as the consultation proceeds, this sort of "ignorance range" would decrease until when you know everything one perhaps could feel very happy about giving a point probability. There are also imprecisions in the measures of uncertainty but one can think about this as the limitations in the knowledge base, rather than the limited evidence on the new case. This range will tend to widen as the consultation proceeds because there will be more and more that imprecise numbers are being used. This is very much an idea of fuzziness, as was pointed out yesterday.

Finally, before I get onto the little slide show, I would like to talk about something that was not really mentioned much yesterday, the requirement for operational meaning of the quantifications. This is something that is not generally given much credence within AI. I feel it is vital that there is operational meaning on three levels of the working of the system.

First of all, the inputs for the system: if there is going to be quantified uncertainty, why should these inputs actually mean something? First of all, it provides a mechanism for agreement among different people who want to contribute numbers to that system, if these numbers actually mean something. They can argue about them and they will argue

about them, but at least they know what they are arguing about. It also provides a means of learning and updating the system. It was asked yesterday how do you update fuzzy numbers as more information comes in. With an external objective meaning or interpretation then they can be updated.

The second reason for wanting operational meaning is the internal coherence, the manipulations within the system as data comes in. It should be subject to scrutiny. It should be explainable to people and be justifiable.

The third reason, which again was not mentioned much yesterday, is about whether the outputs of the system should have external validity and particularly in terms of probabilities the idea of calibration comes in very strongly. If the system is going to be able to sway people to believe what it says, then the output should have operational meaning.

So at this point, I would like to go to a description of a system that I have been working on. I am going to give three examples this morning of systems. The first will be in some detail and the other two will be very brief. This is a system I have been working on for some time developed jointly with a gastroenterological unit in Glasgow . The aim of this study, first of all, was to define symptoms and diseases carefully and collect data, and work out discriminating systems to be able to give some probabilistic diagnosis for new patients.

As I will show in a minute, one of the main aspects of the study is collecting data from the patients by direct computer interviewing. The aim is to identify particularly high risk or low risk groups in order to avoid unnecessary investigations, to save money in the health service. So the system is designed to interview the patient by computer, make some sort of probabilistic diagnosis. That is the bit I want to talk about today, then to make some tenuous recommendations about management and to make some sort of report to the clinician.

But it is the second aspect I would like to concentrate on. The idea is that it can be used on that patient in the clinic by junior doctors, in health centers by general practitioners and in remote areas by paramedical staff. At the moment it is being used in two outpatient clinics and a health center on a very experimental basis.

So the aim is when the doctor sees the patient, the patient should have first of all been interviewed by the computer. The report of the interview does not go to the patient, but the doctor then can get details of the symptoms, an indication of important findings, some sort of probabilistic diagnosis, some suggestions on future management, and the aim is that that should save time in both his interview with the patient and be able to concentrate on important features and in further investigations be able to save money.

It was considered in this that it was very important that explanation facilities were available and we did not just give out blank probabilities of diseases. One of the things that characterize this as a statistical system was that it was based on data analysis rather than subjective probabilities. There was subjective knowledge that goes into the structure of the system but not into the quantifications. So the poor doctors had to do 1200 interviews of patients -- this is just the first sheet of a ten-page form and one can see why this is not done very often. But it gives an enormous amount of information in order to make some sort of developmental diagnostic system.

But now this is one of the original interviewing systems based on a terminal to a DEC 11, with a patient sitting in a booth actually typing away on a special keyboard. The sort of questions that come up on the screen are written in pretty colloquial language, written by a psychologist working with the team. So "does the pain wake you up at night?" "Does this happen often?" "Yes." "When it wakes you up, do you have a little drink?" "Does it relieve the pain?" And a yes to that is indicative of a peptic ulcer. Lots of people wake up at night with pain but getting relief from a glass of milk or a snack is indicative of a peptic ulcer.

I don't think this is a good example chosen there from the interview, because I think that last question is slightly ambiguous, "does this happen often?" It concerns when you wake up at night do you often get relief and I don't think it is quite clear from that question. But the interview is slightly intelligent. It asks about the main symptoms the patient is complaining about first and it just branches depending on the answer. But it is not particularly intelligent. It takes about 25 minutes to go through.

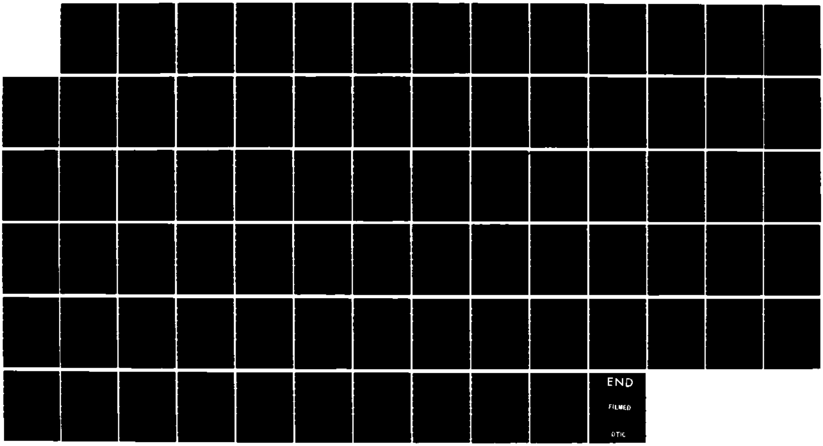
It is now being put on the Apple with a special keyboard. The printer is there by the Apple for demonstration purposes. But it is very easy to use. There is a close up of the keyboard. There is a "don't understand" button which generates more explanation.

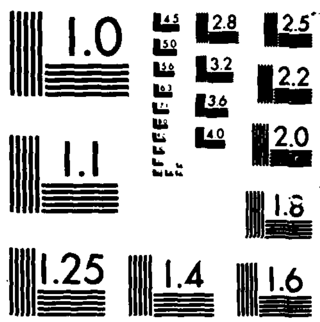
There are six buttons which qualify the degree of certainty the patient has about the finding. These are a complete sham. Any pressing on four, five or six is a yes response and any pressing on one, two or three is a no response. Those buttons were put there because of people saying they want to qualify their answers. I would not mind some suggestions on how those should be incorporated into the analysis. At the moment they're ignored.

This is in Swedish, and the interview is being translated into Dutch and Swedish. The Swedes have written a program so at the end of the interview it can generate a complete letter to the general practitioner about the patient.

So that in itself is a valuable thing, the taking of the interview. It has been shown that people are more honest to computers about how much they drink. People like the computer. Now this generates a vast amount of data. The aim is to produce a simple, accurate device relating the symptoms from the interview to the possible

AD-A163 642 THE CALCULUS OF UNCERTAINTY IN ARTIFICIAL INTELLIGENCE 3/3
AND EXPERT SYSTEMS..(U) GEORGE WASHINGTON UNIV
WASHINGTON DC INST FOR RELIABILITY AND..
UNCLASSIFIED N D SINGPURWALLA ET AL. 15 JAN 86 F/G 9/2 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

diagnosis.

One problem is that people have got more than one disease very often and also there are a large number of questions that are asked. The technique I have adopted in the analysis is directly stolen from many AI applications, which divides up as a "Binary task formulation;" to go through each disease in turn and say what is the probability "they have got it" against "they have not got it." So every disease is considered as a separate task. This is probably not optimal, but it makes it very simple to use and explain to people.

The first thing is to collect single discrimination variables and this is essentially exactly the approach of the abdominal system for abdominal pain and the PROSPECTOR system where you just look at likelihood ratios. Your initial odds on a peptic ulcer would be 194 to 358. Then they say they often wake at night and get relief from their pain by a drink or a snack and we can see that turns the posterior odds, just considering that single piece of information, into 81 to 42. Thus you multiply the prior odds to the posterior odds by this factor in between-the likelihood ratio-take logs to turn it into a summation, multiply by 100 to turn it into a whole number and you end up with what is known as the weight of evidence, which is a term used by Jack Good, which is just the log likelihood ratio.

So what this does is turn an Idiot's Bayes system into a scoring system.

The next thing is to say that Idiot's Bayes is crummy, because it assumes all pieces of information are independent within the disease and the not disease class, so we want to allow some dependence, so you throw this into a logistic regression package and that will tend to squeeze down the scores, what I call crude scores, to adjust them to allow for the dependence between them.

The aim is to produce a scoring system where the outputs actually are calibrated numbers. I will come back to that.

DR. LINDLEY: I don't understand that.

DR. SPIEGELHALTER: What I have done is put the crude scores, which are these crude weights of evidence, and taken those as the data, put them into a regression package.

DR. SHAFER: What is the dependent variable?

DR. SPIEGELHALTER: The dependent variable in the logistic regression is the log odds on peptic ulcer being present.

DR. REEVES: AT no point have you mentioned the physical evidence, like blood tests, urinalysis.

DR. SPIEGELHALTER: I am sorry, I should have explained that. I have not put in the blood test. They have been done on everybody, but this is designed to get the maximum information from symptomatology alone.

And then you can turn the final score, which you get by adding up all of these numbers into a probability. The point being about all this is that you end up with what is a simple system to explain and to use. It makes the computer program actually trivial. As the information is typed in by the patient in response to questions on the screen the numbers are added up inside the machine as evidence towards the disease peptic ulcer. (see slide 5)

What this allows is explanation facilities for the type shown here. At the end of an interview one can state what is the evidence against a particular hypothesis and what is the evidence for a particular hypothesis. This is trying to introduce ideas from AI and put them into a statistical, probabilistic system. So one can say to the doctor what are the important pieces of information and that could perhaps be more carefully checked on the patient.

So one works out the evidence for, evidence against and one can calculate the balance of evidence and in this case it is quite strongly in favor of peptic ulcer. There is an initial score which reflects the prior probability of peptic ulcer in that particular group of patients and that gives you the final score. And that final score can be translated into the probability of a peptic ulcer. (see slide 6)

There is light of conflict which we can introduce, which we are defining at the moment, and I am not sure about this, as being the ratio of the total evidence, that 118 plus 278, to the balance of evidence which is 160. So there is a conflict of 2.5. The conflict ratio would be large if you have got lots of evidence pointing in each direction. It will go down to one if all of the evidence is in one direction.

John Tukey has suggested the output for this program could be displayed graphically, by showing how the scores change starting at the bottom with the prevalent score (on that graph it is shown as a probability) and then the evidence against the hypothesis of peptic ulcer shows it shifting to the left and the evidence for it shows it shifting to the right. There is a graphical representation of the contributing aspects of evidence going up to the final probability on peptic ulcer. (see slide 7)

Here is an example showing conflict in action -- actually not showing conflict. This is for alcohol induced dyspepsia, which is pretty common in Glasgow. These are the important questions about nausea before breakfast, retching, and the point is that the system, by looking at conflict of evidence can identify someone who has all of the symptoms of alcoholism and yet refuses to admit he is drinking. (see slide 8)

If that patient had said "alcohol intake? No, Doctor, I never touch a drop",that would come up as a very large conflict ratio. Lots of evidence is for and lots is against it. But that should ring bells and let the doctor know that perhaps this patient should be questioned more carefully.

This is an example in the paper evidence for and against gallstones. It is a sort of an account sheet of evidence looking at conflict, and justifying the final chance. You notice I say chance there. There is a good reason why I am using the word chance, because of what we actually mean.

What I would like to talk a bit about now is the idea I mentioned earlier about ignorance. How does one incorporate ignorance into a probabilistic system? Ignorance, I view as meaning it is very feasible that the probability that you have at the moment could change very dramatically, because it is based on very little information. One way that one can look at this is to say what are the possible probabilities that can be obtained by a patient at a particular point in the interview.

So before an interview starts there is the distribution on the possible scores that can be obtained for diagnosis of gallstones. Most people are going to get very low scores; some will get fairly high scores and the ones shaded are the people actually with gallstones. So what we consider when we start an interview, the probability of gallstones is only about five percent. But that could change dramatically as information comes into the system, so we would say that 5% probability was based on considerable ignorance.

Why I called those probabilities chances is that because of the way in which this has been designed, the data analysis that has gone into the system, these probabilities are calibrated. When the system says 61 percent chance of gallstones, then round about 60 percent of the time it is going to be right. This shows a rough calibration curve where you plot along the bottom the probability given to a disease by the computer system and along the side the actual proportion of times the disease actually turns out to be present.

If the probabilities mean something, if they can be calibrated then that line is about on the diagonal. The solid line are the doctors and that is a typical pattern of gross over confidence. They say "I'm 99 percent sure it is a peptic ulcer" and they are only right about 80 percent of the time. They have gotten better now with training. So the point is you can go through this and one gets probabilities. They can add up to more than one, because you have got multiple diagnosis. They could add up to a lot less than one.

I don't really want to emphasize actually making recommendations. Essentially for low probability, we would say you can ignore the disease. For very high probability, you should investigate it. You should perhaps treat straightaway and in between you should recommend investigations. The point being is to cut down the number of unnecessary negative investigations.

There are some particular aspects that I would like to just emphasize again. The first thing is we have actually coped with hierarchical diseases, although that did not come out in the presentation. Dyspepsia is a hierarchical disease structure in which, for example, the disease peptic ulcer breaks down into duodenal or gastric ulcer and many symptoms come in at the level of discriminating peptic ulcer from people who have not got a peptic ulcer. A few symptoms come in for discriminating duodenal from gastric ulcer.

How we deal with that is to treat those as two separate problems, essentially in the hierarchical taxonomy to do one division between peptic ulcer and non-peptic ulcer and get a probability of peptic ulcer and then get a probability of duodenal ulcer, given it is within the peptic ulcer class. That comes out from a separate scoring system. These can be combined essentially by multiplying the probability down the tree to give an overall probability of duodenal ulcer.

So within this fairly simple taxonomic structure, we can handle it using probabilities. Other aspects concern the idea of probability ranges. Before the interview starts, the probability of gallstones is about five percent and it is a fairly tight standard error around that value. That is an imprecision coming from a knowledge base. So we start off with a fairly precise probability, but one that is almost totally vacuous for decision making. You can say probability of five percent but it is based on almost no evidence whatsoever. That is reflected by the fact that you know at the end of the interview the probabilities can range anywhere along this distribution.

So your predicted distribution of the final probability that could be taken on is very wide.

DR. WISE: Is this taken to be a data distribution?

DR. SPIEGELHALTER: Yes, this is just an empirical distribution of the final probabilities that it could take on. After the interview is finished, say, in that patient that we talked about earlier, you get a probability of gallstones of 61 percent and that has actually got quite a wide standard error around it. This can be calculated if necessary. So the probability 95 percent interval is 40 percent to 78 percent, so it is a big, fairly imprecise number, because you have put in a lot of these scores with error attached to them. So you end up with imprecision in the number multiplying up as the consultation proceeds.

DR. SINGPURWALLA: You assumed independence.

DR. SPIEGELHALTER: Not quite, because of the regression analysis. So it is a fairly imprecise number but one based on considerable evidence, compared with the initial number which is precise but pretty vacuous and ignorant. And I will come back to that in a little bit, but I better just move on.

I would like to talk very briefly now about another system some colleagues have been working on. IMMEDIATE (Intelligent Modular Medical Information for Assessment, Treatment and Education). This is a real prolog based expert system, rule based. It looks a bit like PROSPECTOR and it is designed for general practice. This is a particular module of it dealing with gynecological problems.

I just want to talk about one particular aspect which relates again back to the idea of ignorance and possible probabilities that can be taken on in the future. The important thing about a computer system that is going to be used by general practice is that it should be very unobtrusive and should only ask questions when it is convinced the question should be asked. It uses this idea of importance driven control, where new information comes into the system and uncertainty is propagated through the nodes, using what we are attempting to have as a coherent calculus, but we are struggling a bit on that one, and then backwards come some idea of importance of questions that have not been yet asked. This suggests an ordering of questions that can be asked by the general practitioner that come up on the screen indicating their importance to be asked.

So the idea of importance is very important. One can think of it as sort of an ad hoc way of trying to do a decision analysis, to try to ask what are the important questions to ask next.

DR. WISE: You say you are planning on revising this ordering?

DR. SPIEGELHALTER: Yes. I can't explain the computation techniques, but it is pretty heavy stuff, because you have to do a complete search forwards, the whole time. What goes into deciding on whether a question should be asked or not has to do with the current certainty of the question.

And the "certainty limits." These are the possible extreme values that that probability could take on when further questions are asked. This summarizes our current ignorance about that particular question, ignorance specifically related to what we don't know but could know within the system. There is also a measure of the potential importance of the answer. And you end up -- I mean this is an amazing phrase and I am not responsible for it, an "importance actualization function," which combines these three components and in an ad hoc way, is trying to get over an idea of the expected change in utility, to give an idea of investigative importance.

DR. DEMPSTER: David, are all of these things formulas?

DR. SPIEGELHALTER: Yes.

DR. DEMPSTER: So there is some mathematical function?

DR. SPIEGELHALTER: Yes, but it is completely ad hoc.

DR. DEMPSTER: It is not totally fuzzy in terms of words?

DR. SPIEGELHALTER: No. Again, I would like to come back to this idea of formalizing ignorance, formalizing what we don't know, which relates both to the dyspepsia system and to IMMEDIATE. Let X be what we currently don't know within the system, the nodes that have not been established yet, so these are knowable things but yet haven't been asked. Let P be our current belief in some disease D. Suppose that were we to observe little x, we'd end up with a final posterior probability of the disease D. What we should try to do is calculate the predictive distribution of the final probabilities that could occur, and use this for control and explanation purposes.

Now, what I am saying is that one has a certainty of any hypothesis at any time, but the idea of ignorance is interpreted as meaning that it could change dramatically when new information becomes available and that is actually formalized by carrying over a particular distribution on the possible probabilities that could occur when more information comes in. However, IMMEDIATE only looks at the range of these, but in the dyspepsia system we are trying to incorporate the entire distribution.

Now, with a trivial bit of sums, we actually find that the expectation of this distribution of final possible probabilities is in fact the initial probability. So all we can say is our current belief in any hypothesis can be looked upon as the expectation of the future belief that we might have when we finally finish the consultation.

DR. SINGPURWALLA: How did you average out the evidence X in so doing? Isn't X the evidence?

DR. SPIEGELHALTER: X is the stuff we don't know yet. It is the questions we have not asked yet.

DR. SINGPURWALLA: So you are averaging out?

MR. SPIEGELHALTER: We are averaging out with respect to what we haven't asked yet and we end up with just what we know at the moment. It is a reasonable thing. It is just a martingale.

So this is the idea of a distribution which reflects our current ignorance, which narrows as the consultation proceeds, because the final posterior probabilities are narrowed down further and further. Now, in general, this definition of ignorance in terms of explicitly what we don't know yet is not possible within general statistics, because one can't enumerate all of the questions that have not been asked.

However, what characterizes an expert system is it is a closed body of knowledge. The actual computing is heavy, because one has to work out all of the time what has not been asked yet. But one can theoretically work out a predicted distribution over all possible answers that could occur when the consultation is finished, and that

provides a formal definition of our current ignorance concerning the truth of our hypothesis. That is only possible within closed systems, such as expert systems.

I better just be drawing to a close now. I would like to go on to talk about relation to fuzziness. I have talked completely within a probabilistic framework and I believe there are areas where probabilities won't work, and I will just very briefly talk about those now.

When do probabilities make sense and when don't they make sense? The first thing is if you do have the idea of the degree of truth, if your propositions are not crisply defined, then it seems quite reasonable to use some type of fuzzy measure. However, I believe that one can often avoid this in our computer interviewing system. We might have questions that might appear not to be crisply defined, such as "Do you often wake up at night?"

Now, the statement, "the patient often wakes up at night," one can think of as a pretty fuzzy statement. But if one only interprets the "patient often wakes up at night" in terms of "when asked whether he wakes up often at night, the patient has pressed the button yes" and if the statement "the patient often wakes up at night" and the explanation and all thinking about the problem is always viewed in those terms, then that is a sort of cheap way to crispify the statement. The phrase should not even be interpreted as being the truth about the patient, but should only be interpreted in terms of the specific button that has been pushed when the person has been sitting in front of the expert system. So that seems a way around some of the doubts about fuzziness and how propositions can in fact be crisp if they are given that interpretation.

The other way our probabilities might not make sense is if the propositions are not verifiable. This would seem to be most appropriate within the control of systems. I mentioned yesterday in statistical expert systems you might have got unverifiable propositions like assumptions one would like to make, like there are normal errors and there are linear relationships, which are useful for control. They are essentially conclusions, interim conclusions one would like to make.

It seems quite reasonable that some other calculus might be used in order to justify those assumptions, rather than probabilities. Also for control purposes there may be an idea of a degree of support for conclusions, a compatibility of one set of data and one set of hypothesis.

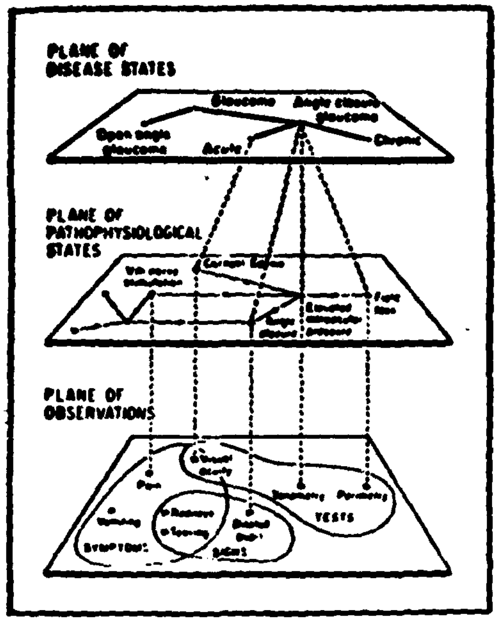
I better stop now. What I have argued about is that in dealing with uncertainty, it is not cut and dry. There are many linguistic ways in which it is used and there are areas where probability might not be the appropriate thing to use, in particular for control purposes in drawing conclusions.

However, probability I feel is enormously more flexible than the way it has been portrayed. I think it can cope with hierarchical disease structures. I think it can cope with conflicting evidence and explanation in the way GLADYS does and I believe that it can cope with some degree of imprecision, although I am still not quite sure of the best way to do this. In particular, I believe it can cope with the concept of ignorance within an expert system when you can specifically state what you don't know yet. You can say the potential probabilities that can be taken on and that can be used as a description of one's current ignorance.

But the major advantage in probabilities is, I believe, that of operational meaning. The inputs mean something. They can be argued about. The manipulations mean something, although as I pointed out yesterday, trying to get a network system to propagate probabilities in a coherent way is a very difficult problem. And the outputs can be made to mean something in terms of their calibration and their interpretation for future action, and I feel that that is the argument that I find convincing. And I must say in discussing it with my clinical colleagues they also find it convincing.

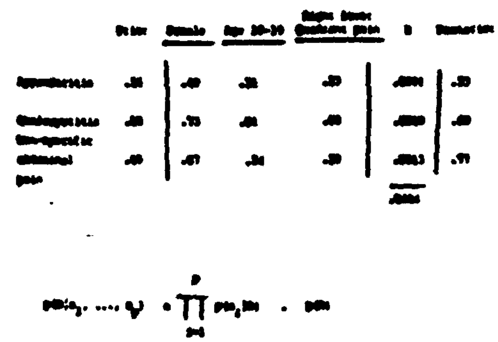
Thank you very much.

(Applause)



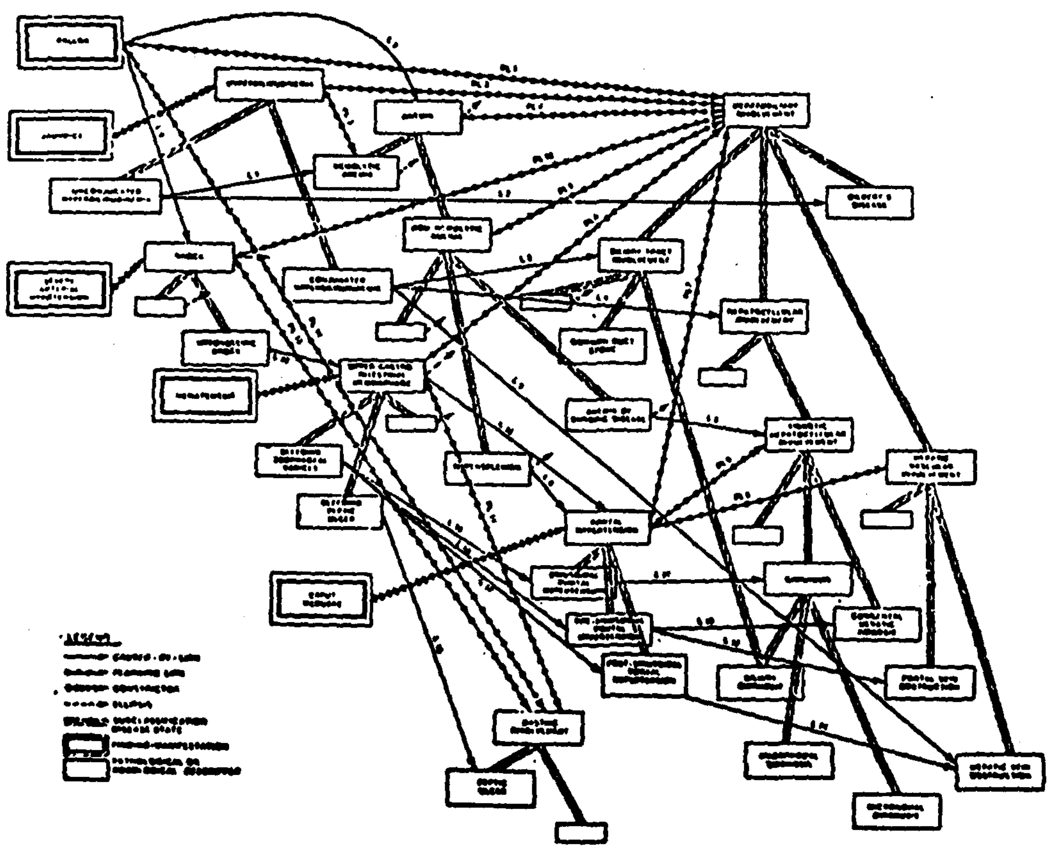
An example of the three-level description of disease abstracted from the CASNET/Glaucoma model

Slide 1



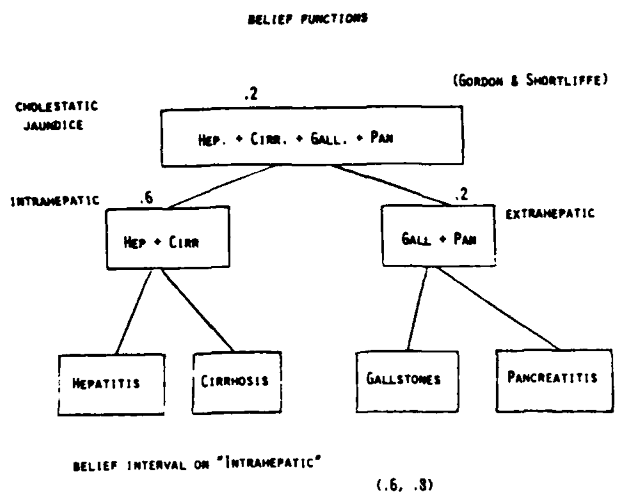
IDIOT'S BAYES

Slide 3



CADUCEUS knowledge structure

Slide 2



Slide 4

SYMPTOM SCORES - PEPTIC ULCER

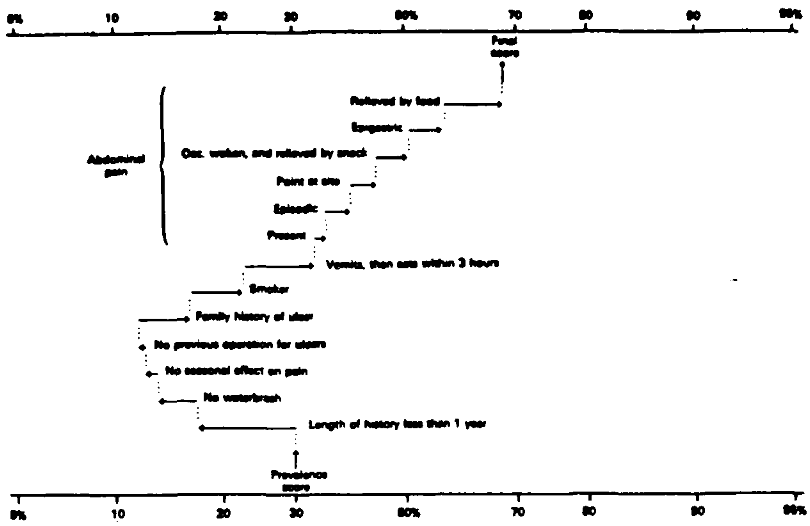
Starting score -84

Indicant	Present		Absent
Pain in epigastrium	28		-50
Pain comes in episodes	19		-57
Pain worse in winter	91		-9
Wake and relief often	100	25	-50
Length of history >4 yrs	69	-3	-75
Previous ulcer operation	125		-5
Family history of ulcer	39		-26
Smoker	41		-74
Pointing sign	56	19	-18
Relief from food	44		-42

Slide 5

Evidence FOR Peptic Ulcer		*Evidence AGAINST Peptic Ulcer*	
Abdominal pain	(+9)	Length of history less than 1 year	(−75)
Episodic	(+19)	No previous operation for ulcer	(−5)
Relieved by food	(+44)	No seasonal effect on pain	(−9)
Occasionally woken at night and relieved by snack	(+25)	No waterbrash	(−29)
Epigastric	(+28)		
Point at site of pain with fingers	(+19)		
Family history of ulcer	(+39)		
Smoker	(+41)		
Vomits, then eats within 3 hours	(+54)		
	+278		−118
Balance of evidence	+160	(Total evidence 396: conflict ratio = 2.5)	
Initial score	−84	(corresponding to prevalence of 30%)	
Final score	+76	= 68% chance of peptic ulcer	

Slide 6



To be used for distribution of final score for patients with peptic ulcer

Slide 7

SYMPTOM SCORES - ALCOHOL INDUCED DYSPEPSIA

Indicant	Score	Cumulative	Probability
		-444	
Male	104	-340	.03
Single/separated	70	-270	.06
No abdominal pain	129	-141	.20
Nausea before breakfast	126	-15	.46
Retching	75	60	.65
Painless diarrhoea	91	151	.82
Heavy smoker	81	232	.91
Alcohol intake - heavy	373	665	.99

Slide 8

DISCUSSION ON PRESENTATION OF DAVID SPIEGELHALTER

DR. DeGROOT: Thank you very much, David. I guess we will follow the same format as we did yesterday. I will give Art Dempster a chance to comment, if you have comments.

DR. DEMPSTER: Actually, I have been operating pretty much in the learning mode for the last hour and enjoying myself. Of two things I might mention, one is from a Bayesian perspective. A Bayesian perspective to me more or less includes belief functions. That they are the same kind of thing is a reaction to GLADYS.

GLADYS, from reading the JRSS paper by David and Knill-Jones, was based on data from 1200 patients, I believe. So it is really a statistical system. The kind of question I wonder about, after thinking about the AI approach to things, is wouldn't you have been able to create a pretty good system without ever using those 1200 patients? From a Bayesian perspective, then, that is using a prior.

Why has all of this prior information been left out of the picture totally and could not one do twice as good, and whatever, if you used it?

The second comment on a totally different topic has to do with the narrower issue, this business of tree structures on the diseases that Glenn and David mentioned. I was just going to mention that Augustine Kong's thesis develops models in the belief function framework, which I think can be quite useful for pursuing that kind of thing. If the chairman wishes, I am sure Augustine could tell us about that for five or ten minutes.

DR. DeGROOT: I call on Stephen Watson.

DR. WATSON: I too found this talk very intriguing and very interesting. It is nice to hear someone who has actually spent some time constructing one of these systems. As David was talking, the questions that were rising in my mind were, in constructing one of these systems, that all of the time one needs to make analytical decisions in the process of constructing a model. Modeling decisions, shall we say?

First, in constructing a model, you have to decide just how to do it. Do you do it this way or do you do it that way? One of the theories that that subject does not seem to have very strongly developed is the theory of how to construct models which use some of these ideas. It is really a question of validation. How do you know that some particular system you have constructed is a good one?

Now, I have really no answers to this, except to say that there seems to me to be four different points that are worth making. I would value David's reaction as to whether he actually did use any or all of these four principles in validation in constructing this system. Or if he feels they are used in the construction of a similar system.

First, the faithfulness to a normative principle, for example, probability. David said yesterday with some force that probability is the only way of handling uncertainty and if this is the case and we are constructing an expert system which is supposed to reflect uncertainty, then the question we want to ask of that system is how well is it using probability theory.

And if it is not using it, is it just not using it because the probability theory is too difficult and you are having to find some approximation to it, or is it actually going against the principles of probability theory in a way that is unacceptable?

Calibration is one David did mention. I was interested to see that he had done this. This is something that has come to my mind and I was interested to see that in this particular case one was able to say this model is a well-calibrated assessor of probability. To the extent that you believe in probability it is a good thing many expert systems are not designed to produce a probability. They are designed to make decisions and actually affect control.

Now, in that case, you can't use calibration so obviously as a criterion into validation. But in a system which is designed to produce a probability, you obviously can.

User satisfaction is another criterion we use in validation. I find it a very difficult one to go along with. Because how do we know that the user is right to be satisfied? The danger is that we are selling him flimflam or packaging and we are not selling him anything which actually does anything for him.

There are lots of techniques that people peddle. Sometimes I fear that they are peddling them well and get a lot of user satisfaction because they are good marketing men, not because they have got a good product. So I think user satisfaction is, for validating, something that you must treat very carefully.

I felt that David was actually using that quite a bit in what he was saying about the clinicians needed to have certain qualities in their aim before they would be prepared to use it. The user had to be satisfied they were doing something. Perhaps user satisfaction has to be gone along with only insofar in doing so we don't go against some of the other principles of validation.

I suppose the most obvious one in judging whether an expert system is good is to see if it compares well with expert performance. Now, here again I think there are two views in the literature on expert systems. Either you are trying to construct something that does it well or as best human around, or you say the best humans around don't do it very well and we ought to construct something which is better than the best humans around.

If the latter position is taken you don't want it to be similar to expert performance. You want it to be perhaps faithful to the normative principle. These are some ideas and I would like David's response to them.

DR. DeGROOT: Okay, David, respond.

DR. SPIEGELHALTER: First of all, the thing of why do we use prior information. You are quite right, we nearly broke their hearts back in the states. The project nearly collapsed and I did not come into it until just about the end of their data collection exercise. They ran out of grant money and it was a terrible waste not to have something worked out. I mean they started a long time ago, since so many of the techniques that I have been using are direct ripoffs from AI stuff that I have only just come across, that I don't think we could have built this thing five years ago.

But, yes, certainly the most sensible thing in designing a system like this would be to start off using completely subjective opinion and then update it in proper Bayesian ways as data comes along. We are trying to do that with one, working in chest diseases at Westminister Hospital, with a system for diagnosis for asthma, et cetera, to be used as an initial test to decide about further investigations. And there we have got the clinician assessing quite a large number of subjective probabilities and particular findings, the probability of yellow sputum in a male in this age group, with shortness of breath, and no chest pain, no coughing up blood and asthma.

So we look at fairly restricted disease groups and ask them for these subjective probabilities. Now, quite reasonably he does not mind being very precise about these and as people have said, I think the number if fairly high, well ---

DR. WISE: I am sorry to jump ahead, but when it says 50 to 90 percent, is that one standard deviation or two?

DR. SPIEGELHALTER: In the questioning -- I mean, we generally take it so if it is outside that, he says he would be pretty surprised. We try to get them so we take them as being about one in 20 charts and so we will make them, give an interval for which they would think they are 95 percent sure. Now, whether that 95 percent means 95 is a different subject. So we interpret that, when he gives a range, we interpret that as a 95 percent credible interval, or whatever, based on an imaginary sample. That in fact corresponds to the interval that you would obtain were you to observe an imaginary sample of 20 asthma patients, of whom 14 had the symptoms.

The number is stored in the system, not as a probability but as a fraction. It is stored as 14 out of 20, so that when new data comes along and, say, we observe another ten real patients in this group, nine of whom have got the symptom, then we can update that number, that fraction into a new fraction and so update the probability.

DR. WISE: But if that formula is right, you are implicitly assuming that it is a data distribution.

DR. SPIEGELHALTER: Well, no, there are, you can update using means, sample means. All right, a data distribution would give that, but you can do updating with these combinations of mean, prime means and sample means under much more general assumptions than data distribution, but so who cares? It is just a rough idea.

The joy of doing this so we can actually get a system off the ground within a month's work -- okay, it is a lot of argument; they sat around and bickered about these numbers a lot, but then you actually get them going and you are updating it and that is just the right way to do it.

DR. WISE: When you get these numbers, how many clinic clinicians do you interview?

DR. SPIEGELHALTER: That was just based on a couple of people.

DR. WISE: And you make them give you one range.

DR. SPIEGELHALTER: We did not do a Delphi technique on that, but just sat down and argued about it. So they are pretty crude things to get it off the ground, which we hope that the data is going to be sufficient.

DR. DEMPSTER: On the other hand, David, this trial that you mentioned about the 16,000 people, that should produce wonderful data.

DR. SPIEGELHALTER: No, they refused to put the symptom data. It was all wasted. It was due to the organizations in ten hospitals and all of the stuff is being punched onto micros. The symptom has been punched on, but it is not being kept. So that was a tragic waste.

And just to answer Stephen, I agree user satisfaction is important and our design is changing always as the criticisms are made of it. In terms of the evaluation, whether you are trying to get right or whether you are trying to get better than the next one, the final thing is that the evaluation does not affect patient care. Probably this is most important.

There is now becoming fairly established a four-stage evaluation procedure. It is following almost exactly the same pattern as the user trials, where you start with such as initial safety, and then eventually to stage three direct trials and do a control trial. Each of these things is important and can be tackled as separate evaluation systems.

DR. DeGROOT: Thank you, David. Are there comments, questions from the floor?

DR. BROWNSTON: I was struck by an analogy between what you are doing and something which is quite old fashioned and that is psychological testing. There are a lot of different ways of looking at psychological testing. One type of testing is very much like expert systems, like R-1 or Maxima or Dendral, which are quite deterministic. The analogy in testing would be to exhaustively ask the students the multiplication table and see if they know the multiplication table; or give them certain things, like can they integrate trigonometric functions by giving them trigonometric functions integrated.

In that case there is really no uncertainty involved and you get a deterministic answer. Another step up would be in personnel testing or aptitude testing, where your psychological test is somewhat like an expert system to do an interview. You ask them a sample of questions and on the basis of the answers to these questions you determine whether this person should get this job, or should get admitted to a university, and so on.

There has to be technology for determining whether this psychological test is doing what you expect it to do. There has to be external validation. So you have to do follow-up studies to determine whether upon using this test you get a higher proportion of successes in the admission procedure than if you did not use this test. This is what is called validation.

But there are other techniques for determining coherence which are called reliability tests. In this case it's the validation which is similar to calibration, in the sense you are trying to maximize the number of correct decisions. Then there is also actually an analogy to user satisfaction, and as Stephen Watson pointed out, it is called based validity, which is considered to be public relations in testing.

You have to put in enough questions to make it look like you are testing what you are supposed to be testing, even if you can determine a person's success in college by asking them if they like eating raw carrots. There is a test which asks irrelevant questions like that, which happens to be valid.

So this brings up all sorts of interesting questions about the nature of building expert systems, especially about validation. I think validation is one of the most important things. When you can validate, you must validate to determine if your system is doing that. Then after these two uses of tests, there is a third use of psychological tests and that is personality testing. Where you are not even sure whether introversion, extroversion is really a valid dimension of personality, but you are using the test in an exploratory fashion.

Just so in some domains the question of validation is very difficult. One of them would be military decision making, because you don't have a set of observed frequencies. You only have expert judgments that you can use. So what your expert system is doing is trying to simulate what the expert knows. In fact, it is perhaps even doing psychology in trying to do that.

So this is more fuzzy. The progression I was giving was something very deterministic to something probabilistic to something which we really don't know what we are doing. And I think that progression goes into both psychological testing and in what you are doing. I am glad to see that you are solving problems in very similar ways to the way that people in psychological measurement can solve them.

DR. SPIEGELHALTER: Yes, the parallel is incomplete there, and we are interested in working there. There are exhaustive -- and in a sense this can be viewed as a way of trying to avoid doing exhaustive deterministic testing, you could find out what is wrong with them. This is trying to avoid doing that, instead of asking people questions, irrelevant questions, you ask a few of them and try to judge essentially how they would have answered the rest.

The final area about the ill-defined central final outcomes is very widespread and something I try to avoid. Our disease categories are well defined, but is is very complicated.

DR. ZADEH: I found this to be a very impressive piece of work even though there are a great deal of ad hoc procedures of one kind or another. I think these procedures are basically unavoidable. In other words, you just can't use formal hearing in this sort of application.

I think there is one problem with approaches of this kind and that is that you might get drowned in an ocean of information. For example, we have this ten-page questionnaire. The problem is, with things of this kind, that anyone of the answers in that questionnaire by itself will not be decisive. It is a little bit like the following: Suppose you want to decide on whether or not to promote an assistant professor to associate professor. Instead of asking a few key people, you ask the students and the secretaries and people not in the department and you get 10,000 opinions as to whether the man should be promoted or not. Well, the 10,000 opinions of that kind when aggregated together would be much less reliable here than the few key assistants.

One of the problems that plays an important role is the issue of control strategy. What question do you ask next, because depending on the hypothesis that you are sort of converging on, a set of questions may or may not be relevent. The uncertainty in some of these systems, the more sophisticated ones, is used to determine what question to ask next. This is the way a doctor would proceed.

The way a doctor would proceed would be very much influenced by a tentative hypothesis that is being formed in the doctor's mind and the perception of what questions would be central to that hypothesis.

There was a little bit of mention of that thing in the branching questionnaire. In the case of this last thing when you said, when you try to define ignorance, you say, okay, we will try to compute the probability as a function of the possible answer to the question. To me this is totally an impossible enterprise. In other words, it would be a mind boggling exercise. Even in the case of closed systems you would qualify that. The possibility of different questions to different

answers and the impact that this may have on probabilities would be such the answer necessarily would be that the probability would be zero and one and the expected value is .5. It seems to me that no other answer really could be obtained to a question like that.

In other words, if you have the wrong questionnaire and you are trying to compute the probabilities as a function of the possible answers to this, I don't see anything else in there. I don't know if you have actually done it or not, but it seems to me that if you have not done that and if you would do it, this is what you would be arriving at in total ignorance.

DR. DeGROOT: David, do you want to respond?

DR. SPIEGELHALTER: Yes, there are a number of things. First of all, in GLADYS we consider the patient's time is free. It is costless and so they have to answer a lot. They get set down in front of the screen and have to sit there for a half hour or something like that. So there is not the need for the control strategy.

In the other system I was talking about, the one for general practice, then it is a very limited number of questions and then there is the need for a very stringent control and identifying very important questions, so that should be incorporated. The idea of can one actually do the search and work out the possible values that a probability could take on, this one is quite reasonable that if you ask enough questions you will get as close as you need to zero and one. However, if one can do actual distribution and you don't pull in on the range, then still that distribution can tell you what your ignorance is about a particular question. It shows you how sensitive your current belief is to further information that you could obtain.

The actual computational difficulties are difficult and this is what I am working on with these people in IMMEDIATE. They think they can do the search through in order to generate at least range. And clearly if you generate the range it is zero to one but it is useless, as you say. In which case you really want to get the whole distribution. It is reasonable that any state in an expert system can generate plausible further findings it is going to have.

DR. YAGER: I just want to say something about the validation. I agree it is a very important issue. It seems to me that there are at least two considerations one has to have. One is how truthful the system is, how good it predicts the right answer.

For example, if you have a weather forecast, in an expert system it sort of predicts a high temperature tomorrow. So one consideration is the issue of whether it does indeed predict the right temperature tomorrow. But a second consideration, a very, very important one, is how specific your answer was in the sense that if you have an expert system that predicts that the weather will be, let's say over 30 degrees tomorrow, that is his prediction, it is very unspecific. It will always be correct but it won't be that informative.

So I think for validation you have to consider two factors in the sense they contradict each other. The correctness of the answer as well as the specificity of the answer, if you want a specific answer. If the answer is not very specific, even though it is correct, it is not useful. I think you have to always consider those two conditions and they sort of fight with each other.

DR. SPIEGELHALTER: That can be done formally. If your prediction is done in terms of distribution, then one can use a scoring rule that is related to the order of that distribution at the true value, and so it is way out in the tail. It may be a value you have given some support to, but it is way out in the tail and there are lots of other values you consider much more likely. So it is pretty useless.

DR. DeGROOT: It is time for a coffee break. I thank you again, David.

(Recess)

DR. DeGROOT: Let's resume the discussion that we were having. If there are questions pertaining to David Spiegelhalter's talk, we will welcome them now. Also, it would be nice to hear from others who have had experience with particular expert systems. I would be very interested in hearing experience as to the practicality of using the different methods that we have been discussing in real live expert systems, the practically of implementing Bayesian methods or fuzzy methods or belief function methods and so on.

If you have such comments, experiences, even if they are not tied directly to questions or the particular talks, I am sure there are many of us here who would like to hear about those, so don't be bashful about telling them to us.

DR. SINGPURWALLA: This is really not a question, but more of a comment. It seems to be coming up over and over again and I think it pertains to this hierarchy that David showed this morning. I am constantly reminded of fault trees and event trees. The important point that I want to make is that you mentioned the notion of importance somewhere along the way, and you said it was wavy, heuristic, and so forth.

In fault tree analysis, we do have the notion of importance, which is mathematically precise. There are two measures of importance. One is what we call structural importance and the other is what we call reliability importance; the latter is a probabilistic notion. I am suggesting that these kind of notions be considered in the context that you are interested in, and you may find them useful.

You can look at a certain node and judge its importance based on its structural position and also based on the probabilities that you are willing to assign to it. I believe Professor Zadeh was also referring to the question of importance.

The second point is that somewhere you mentioned the inadequacy of probability theory. You cited two situations. One pertains to outcomes which are non-binary. Is that right?

DR. SPIEGELHALTER: I meant just ill-defined propositions.

DR. SINGPURWALLA: All right, I won't pursue the non-binary issue for now. The second issue you mentioned were questions of linearity, normality of form and things like that. I believe that models are personal expressions and the normality, if ever, in a linear model is something that you as the modeler subjectively specify; and there is nothing about its truth or falsity.

DR. SPIEGELHALTER: Yes, it is an idea of assuming something, but in order to assume normality within a system one wants to get some idea, is there evidence against it, which suggests some sort of known probability.

I would not like to talk about the probability of normality. I don't really feel like I knew what I was talking about at that point and so I am prepared to see that there may be areas in terms of control strategies where you do want to make assumptions where some strictly non-probabilistic measures have evidential support or something might be used.

DR. SINGPURWALLA: Getting back to this issue of normality and linearity of errors in linear models, the sample theory approach to these issues would be the analysis of residues, and wouldn't that still be within the framework of probability?

DR. SPIEGELHALTER: No, it would be in the framework of tail areas which is not in the strict probabilistic range. You can make some judgment based on the tail area in the distribution consisting of hypothesis, which is not saying anything about the probability of the hypothesis. So that is, strictly speaking, nonprobabilistic reasoning.

DR. SINGPURWALLA: The third comment I think is a more general one and I don't know what the answer is, but some mathematicians in this audience could probably answer it a little bit more precisely. How much of our knowledge of mathematics is based on the notion of binary variables?

If much of it is, then arguments against the use of probability theory essentially would not in any way fill that gap. Everything that we can think about is in binary terms and I would think probability theory is adequate to deal with it.

DR. SPIEGELHALTER: Again, it is ideas of imprecision of the fact that something could be probably true.

DR. YAGER: It seems to me it is sort of a moot question in a way, because if everything relates on binary then all of arithmetic -- I mean then why do any arithmetic other than with ones and zeros. Why introduce the numbers three, five, seven and so forth? Maybe you can do everything, I am not sure, but maybe you can do everything in binary.

But the point of the matter is that is not the most effective way to think, or to communicate, or to do manipulations. So I don't think it really matters.

DR. SINGPURWALLA: Three, four is a build up on the binary system, so I don't think your counter example is that good.

DR. WISE: It seems to me that your objective is better met by the example of real numbers, which they have not even, their cardinality does not even correspond to binary things. But you can talk about real numbers with propositions which are themselves either true or false. Every formula you write, plus a set theory, is a proposition which is either true or false, when you talk about the probability that a real number falls in an interval, or in another real interval, or another. So in one way you are handling continuous things very easily but you are working only those propositions which are true or false and you are talking about probability theory.

DR. SINGPURWALLA: You have not told me that there is a multi-valued logic.

DR. WISE: There are those too, but those are the propositions.

DR. SINGPURWALLA: You just have two.

DR. WISE: But those are propositions. In multi-valued logic you have just got different propositions which are themselves either true or false. They have no value. They may have a value true or false or unknown one or unknown two, but they either have that value or they don't have that value.

DR. YAGER: But you could go on ad infinitum and you can add your multi-value logic on.

DR. WISE: Sure, math gets complicated.

DR. ZADEH: There is a more basic issue of that that calls into question this kind of physical analysis. That is most of the events, most of the propositions here, are really fuzzy events. I think you will admit that when you talk about high fever or hardening of the arteries and when you talk about having gallstones. All of these are fuzzy events. In other words, this particular disease could be present to a degree. Now some are more that way than others.

Infectious diseases tend to be sort of yes or no types. Either you have tuberculosis or you don't have it. But degenerative diseases are not like that. They are generally a matter of degree. Even in the case of pain, you have severe pain and you have frequent pain, but how frequent? If you are confronted with a particular situation and you ask and say is it frequent pain or infrequent pain, then it becomes somewhat artificial to force the patient to say yes or no. Because this is not a natural thing.

I think in your presentation you said we sort of treat it as some sort of a proposition, but strictly speaking this is not really valid. It is not really valid and furthermore when it comes to assessing the probabilities then you have to be able to come up with the concept of cardinality. In other words, you count the number of patients who have hardening of the arteries in the presence of certain other conditions. But if hardening of the arteries is a matter of degree, then how do you tell this particular individual who has hardening of the arteries that it is .3 or .7, .9, so that you have a succession of cases but each one of these cases is sort of a matter of degree.

So that strictly speaking none of these probabilities or very few of them can be assessed in classical probability terms, because you are not really dealing with crisp events. So because of this the classical things that we take for granted are not valid. For example, the standard thing that is the case and let's use expert systems in MYCIN pathology (inaudible) is that the conditional probability of A given B is equal to one minus the conditional probability of not A given B. That formula is not valid.

That is if you assume that A and B are fuzzy predicates, if you assume that they correspond to these things like high fever, hardening of the arteries and so forth, that is not valid. So you have an immediate breakdown, an immediate breakdown. You don't have to go far.

The rules of modus ponens break down. Now you try to patch that up, and that is what is done in MYCIN a little bit. You set some sort of a threshold, but these are highly unsatisfactory ways of coming to grips with these issues. What I am trying to say is most people, and that applies to all of us, use whatever techniques they feel comfortable with and they tend to be skeptical of techniques that they are unfamiliar with. This is a very natural sort of a thing.

But at the same time I think that one has to consider the fact that in problems of the order of complexity that David has described, classical probability techniques can be used only in special situations to a limited extent. Beyond that, it becomes a matter of closing your eyes and all sorts of assumptions making by the dozens all sorts of approximations, disregarding dependencies, disregarding the type of things that overlap, disregarding the fact that they don't have sharp boundaries. All of these things are disregarded.

So the question that arises is how much reliance can be put on the numerical probability at the very end, like .25? My contention is very little. It is a label for what is in effect a ball park figure, like low. That is the most that you can say and anything beyond that is unrealistic. True, people use it and it is useful. That does not mean it is not useful. But at the same time we have to have our eyes open.

We have to realize that we will be deluding ourselves if we took those figures seriously, just as the figure of .2 with a probability that Shakespeare wrote Hamlet cannot be taken seriously. It cannot be. So the point that I was trying to make in my own presentation and this is by nature of comment on this thing, it is not a matter of using one technique versus another technique. It is a matter of trying to find an accommodation with a very pervasive imprecision.

In fact, I think in the field of medical diagnosis it is at this point far too complex in relation to the understanding of these issues that we have. We have ventured far beyond what we know about reasoning under uncertainty, imprecision, the issue of what questions to ask, the issue of what tests to perform, considering the fact that these tests have certain risks associated with them and so forth. So that strictly speaking, the level of our knowledge at this point sufficies or experts systems in very narrow, specialized domains in which we don't have the risk of the kind that you have in medical systems.

Now, it does not mean that we should not do that. I think, as I mentioned originally, I was very much impressed by the system that was developed. It is a useful system. It is an effective system. It may have some flaws here and there, but it is a working system.

DR. DeGROOT: What system are you referring to now?

DR. ZADEH: This system that David described. And the same applies to other systems, like MYCIN, so whatever criticisms you make of those systems does not mean that they are not useful systems. It means merely that we cannot merely justify all of these things in terms of formal theories. That is all it means. So that what we could justify at this point, as I said earlier, would be things that would be far less ambitious. That is my feeling.

DR. DeGROOT: Thank you. That was very clearly stated. David, do you have a comment?

DR. SPIEGELHALTER: I agree that if one was genuinely trying to say, was using phrases like the patient sometimes wakes at night with pain and the pain is very severe and from that you are trying to conclude a statement like he has a duodenal ulcer, that is a very fuzzy statement, degrees and degrees of it. Then the calculus of probability perhaps would be unreaonable and you are talking about ill-defined statements.

The argument against that I used before is to say that that is not what we are dealing with. We are in a sense trying to make it amenable to probability and to admit that you have to do this to make it amenable to probability. We are crispifying the statements by saying that everything up there was a shorthand for when forced to say yes or no to the question is the pain -- we don't use phrases like is the pain severe, but when forced to say yes or no to the question do you get early repletion when you eat a meal and the patient answers yes, and from that you conclude the probability that it will be concluded at six months when forced to say yes or no that the patient has a duodenal ulcer, the doctors will say that he has a duodenal ulcer.

So we acknowledge that if we did not make this, put their backs against the wall and make the patient sit down in front of a terminal and make the doctor fill in one box in the form about the final diagnosis, then calculus might be unreasonable and frankly I wouldn't really know what to do, because I wouldn't really know what I was talking about at that point.

But by doing this, by forcing them physically to answer yes or no to a question we are crispifying it. Therefore I feel that the probability is not invalid and that our statements are well-defined and the numbers are well-defined and have an operational meaning. But to justify that one has to see everything that is written down as a shorthand for the statement when forced to answer yes or no to this proposition and they answered yes. It may just seem like a way out, but I think that it does mean that I feel that the numbers we are talking about are valid and are justified and do mean something.

DR. ZADEH: Suppose that a patient comes to you and suppose that you have to check one of these things, the patient has frequent pain. On what basis then would you say that the patient has frequent pain? Suppose he tells you that he has it two times a night or three times a night or five times a night and so forth, at which point would you say that the patient has frequent pain?

What I am trying to say is you cannot sweep this under the carpet by saying that you will treat that as some sort of a crisp proposition, because you will have to indicate which class that falls into. Where will be that threshold? Is it specified or not? This is my question.

DR. SPIEGELHALTER: First of all, we ought to make every effort to make a proposition, the questions as crisply defined as possible. How many times do you wake up with pain and those are often collapsed back down into categories. But that is not really to get rid of the fuzziness solely. It is in order to increase and discriminate repair of the questions.

That I see is the vast advantage of a computer interview for a clinician. Exactly the same question is asked to everybody. There is no judgment on my part. I never fill in the form. Doctors don't fill in the forms. They never have to make a judgment about whether the pain is more frequent or not. It is purely whether the patient pressed the button saying yes or no when asked this question. Now, clearly

different patients will respond differently.

One person may wake up three times a night and one person may wake up once a week and they both press the button frequent. It is quite possible that people have completely different interpretations of the word. But that variation, that respondent variation is taken into account by the discriminating power of that question from the data. So provided that our system is based only on responses to a strict question and that is the way the system is built and that is the way it is supposed to be used, then I don't see any difficulty with the fact that different people may interpret the question in different ways.

If, however, there would be a big difficulty if clinicians just said, oh, this is what I think is frequent pain and put that into the system and then it was used by people who interpreted it in a different way, that would be a very bad thing to do. But provided the term is used with the same amount of vagueness by everybody then it can be treated as a crisp term.

DR. DeGROOT: Glenn, did you have a comment on this point?

DR. SHAFER: I would just like to support on this point about starting the investigation by making things that in the ordinary course of discussion are vague as crisp as possible. It seems to me that is a general aspect of scientific investigation and there is no particular need to apologize for it.

DR. REEVES: One of the criteria suggested was user satisfaction. Do the clinicians understand what they are receiving in way of data is in fact that the person pushed this button, not that the person has frequent urination or whatever it may be and do they accept that as having, as sort of an educational thing as was mentioned before? They say it has face validity, but in fact they are good indicators.

DR. SPIEGELHALTER: Yes, at the moment, because the people using it have been involved peripherally in its development. I think it is fairly clear. There is a big danger though that it could start, what the machine prints out could start being used as truth rather than just what someone responded when asked a question on the screen. At that point, it does start becoming possibly dangerous. The machine's colloquialism is based on a very well-defined, precise idea, which is how the patient pressed the button and I can see there is a danger of the systems in the future.

DR. SOYER: On this problem of validation, for example, normality of errors, I think it is again a problem where you consider scoring. For example, the normality of errors. You never observed errors. You only have estimates of errors in the model, but you will eventually observe whether the patient is going to have a certain disease or not. So when you are validating your model you will be validating based on those observable values.

I think then the scoring comes into the picture again here, because I think for any scoring rule, for strictly scoring rules, it is decomposable into two parts where one takes care of the calibration and the other is a measure of information and I think this can be nicely used in the validation part.

DR. SPIEGELHALTER: Yes, when you have got observable outcomes, then you can use the scoring techniques in decompositions. It is just when you have got unobservable outcomes, it does not seem to be very clear how that has been assessed.

DR. DeGROOT: Let me exercise the moderator's prerogative to follow-up on that with a comment of my own. It seems to me that one big advantage of using probability methods is that it is possible, as you mentioned, David, to do an evaluation of an expert system and to compare different expert systems. In particular one can think about the predicted distribution as you spoke about, the predicted distribution of the outcome or the output of the system of the final probability, for example, of the probability distribution of the various diseases and it is important. Many of you know this, but some may not that when you think about these predicted distributions to evaluate a system, where do I think I will end up after I have collected data on this patient.

We know that we would like our final answer, to use terms that Professor Zadeh and others have been using, we would like our final probability to be as tight as possible. That is, in the terms I was mentioning yesterday, if we think about the final overall probability as the weighted average of various conditional probabilities, we would like in our final answer all of those conditional probabilities to be very similar, then we would feel very certain and reassured that our answer is stable in the sense that a little bit of further knowledge would not change it very much.

But at the beginning of the process, before we have collected the data and we think about what is our final probability likely to be and the uncertainties attached to what that final probability would be, we want that predicted distribution to be as spread out as possible.

The best expert systems are the ones that have as broad a range as possible of where you are going to end up. I mean that is easy to see really, if you think about it, because a useless expert system is one where you know where you would end up before you began, so then you don't need the system, if it is not going to change your prior probability, for example.

So the good expert systems, the refined expert systems are ones where you are very uncertain when you have entered the process where you are going to end up. So the most spread out distributions for your posterior probabiliteis are the best ones. I just raise as a question, and perhaps others will answer it later in the discussion, how that concept of comparing expert systems in those bases could be used outside of the probabilistic methods.

Another way the probability methods enter is in terms of calibration. David mentioned calibration. Good systems should be calibrated in the sense that he mentioned. But calibration as was just commented on, calibration is only part of the story and it is possible to compare well-calibrated systems in terms of what I call refinement or sufficiency or informativeness, that is again you want a system that is not just well-calibrated but one that gives you a wide range of probabilities, as broad a range as possible.

It is very easy to be well-calibrated if you know that 30 percent of the people have duodenal ulcers, then all you have to do whenever a patient comes is say the probability is 30 percent and ignore all of the tests. That system is well-calibrated. It will be right 30 percent of the time and it says it is going to be right 30 percent of the time, but it is useless. It does not take an expert to make that statement.

So what you want is a very sensitive or spread out or refined distribution of probability. So I just wanted to make that point. In terms of probability, it is possible to make the evaluations and comparisons and it is a slightly interesting point if you have not thought about it before that you really want highly variable probabilities. Those are the best systems.

GENERAL DISCUSSION I*

DR. DEGROOT: Okay, let us resume. The title of this session is "General Discussion."

All of the talks are open. Are you going to say something inflammatory to get us going, Nozer?

DR. SINGPURWALLA: I will say a few things as a member of the audience who is used to the calculus of probability, and further enlightened by Dennis Lindley's visit and companionship and at GW.

I am still having a problem in understanding possibility theory, fuzzy logic theory, and also belief functions. And I must tell you quite honestly that I have made several attempts to read the papers.

I have had trouble trying to get at the root of the matter but that may be due to my own weaknesses. The thing that is coming out of today's discussion, applies to fuzzy logic. Please correct me if I am wrong.

It appears to me that fuzzy logic and possibility theory somehow do not apply to statements of uncertainty. I get the impression that it does apply to something which is imprecise; that is, something which is neither yes or no, but something which is maybe, like an item is not failed, but partially failed. It is not raining or dry but slightly raining. Is that correct? I believe I gathered the impression from one of Steve Watson's viewgraphs.

I think he said that fuzzy logic applies to precision rather than uncertainty but one can carry the argument further and say that imprecision implies uncertainty. If that be the case, then my conclusion is that the calculus of probability should be sufficient, the scoring rule argument supporting its basis.

After we settle this issue, I think I would like to ask Professor Shafer and Professor Dempster to make one or two convincing arguments as to why we should be concerned with belief functions and why we should use them. Somehow their message has not come out clearly, at least to me.

(Laughter)

---

*Session followed presentations of Drs. Shafer, Zadeh, and Lindley and ensuing discussions

DR. DeGROOT: I hope we don't have to wait until after we have settled this issue, to use your phrase, before we can discuss some other topics.

(Laughter)

I think it is too much to hope that we are going to settle issues at this conference. I think we are going to expose issues at this conference.

DR. SINGPURWALLA: Who knows? We may even settle it.

DR. DeGROOT: Please, not today, or we won't have anything to do tomorrow if we settle it this afternoon.

(Laughter)

Did you want to comment?

DR. YAGER: I want to comment. I guess I like to use the term uncertainty as sort of a general term and sort of what you call probability. I call that sort of randomness.

DR. SINGPURWALLA: No. By uncertainty, I mean something very precise. What is the temperature outside? I don't know. I will be able to measure it later on. So uncertainty is something which typically reveals itself, unless I am dealing with parameters. Most of the situations of uncertainty eventually reveal themselves. So uncertainty is very clear to me.

DR. YAGER: First of all, in classic probability theory, much of the information is sort of imprecise information. The fact of the matter is that when you have probabilities, for example, you have them imprecisely.

For example, the quantifiers that Professor Zadeh talked about. Most students do this and so forth. That is really a probability. So much of the probabilistic information itself is fuzzy information or imprecise information.

So in dealing with and manipulating probability in the ways that Professor Lindley talked about you have to really manipulate fuzzy numbers and fuzzy information.

DR. SINGPURWALLA: I am not sure I understand you. To me, probability is an expression, a numerical expression of the way I assess an uncertain situation. By definition it cannot be precise. It is the way, one expresses uncertainty about something. It cannot be precise, because it is personal.

DR. YAGER: If I ask you, for example, what is the probability it is going to rain tomorrow, what would you say?

DR. SINGPURWALLA: Oh, I will think about it. I will have background information and based on that I will say the probability that it is going to rain tomorrow is some number P.

DR. YAGER: What is the number you say?

DR. SINGPURWALLA: Well, I will say .8 and I am willing to bet with you eighty cents to the dollar. Eventually somebody will score me on such bets.

DR. YAGER: What I am willing to say is something like, for example, I am willing to allow you to say .8 but I am also willing to allow you to say the probability is close to .8, or near .8, or around .8.

DR. SINGPURWALLA: I may round it off to .8763 if that is what you are after. Because to me the number doesn't mean anything absolute in a certain sense. It is not some physical quantity that I am after. Therefore I fail to see the notion of fuzziness and I am trying to keep an open mind.

DR. DeGROOT: Let me rephrase the question, Nozer. You say that probability is a numerical measure of your uncertainty, but why isn't it legitimate to ask the question "what is your uncertainty about that numerical measure?"

DR. SINGPURWALLA: Your question is legitimate. I could of course add a probability distribution to that original probability. That is, I could add a hierarchy to it, and do everything using the calculus of probability. The concept is very easy, even though its implementation and application might be difficult. I do understand what I would be doing. It seems to have a logical and fundamental basis.

I am at a loss as to why I need these other notions, unless these other notions can make a convincing case.

DR. ZADEH: This question was raised earlier when somebody asked Dennis what is uncertainty. And there is of course an obvious answer to this question. Some people, and I think Stephen took that position, would differentiate between uncertainty and imprecision.

There are other people who would say imprecision is simply a kind of uncertainty so there is sort of a hierarchical relationship between them. I tend to take the latter point of view.

So certainty is something very general and there are different kinds of uncertainty. But when you talk about imprecision or fuzziness, you are talking about the lack of sharp boundaries.

It is sort of a situation where membership in a class is a matter of degree. So when you are talking about somebody being young or whether it will rain tomorrow or not, these are matters of degree, even more so in the case of whether it will be warm tomorrow or not.

So in probability theory, the concept of an event is a crisp concept. In other words, either something is in the event or it is not. No allowance is made for situations in which an event can take place to a degree.

One of the things that fuzzy logic does is that it makes it possible to enrich your language by allowing you to deal with fuzzy events. And furthermore--and this is also an important characteristic--it makes it possible for you to describe the probabilities that are associated with fuzzy or crisp events in imprecise terms.

By so doing then, it gives you more tools to work with. It gives you more a expressive language. So essentially the disagreement is this. There are some people who say that the language of probability theory is sufficient. It is adequate. Professor Lindley is a foremost exponent of that point of view.

There are other people who say no, it is not adequate. It is not a matter of saying that probability is wrong or right. It is a matter of adequacy. So the latter position then is that it cannot cope with problems in which the events are fuzzy events.

It cannot cope with situations in which your characterization of probabilities is imprecise. It cannot cope with those problems. This is really the position that is taken.

So long as you stick to problems in which the events are crisply defined, probabilities are crisply defined, then you stay within your probability theory. There is no problem.

DR. LINDLEY: Well, I want an example. You talk about rain tomorrow. That is perfectly crisply defined in terms of the probability distribution of the millimeters of rain that will fall. Any fuzzy statement that I have heard you make can be stated in probabilistic terms.

DR. ZADEH: What about a warm day, because warm is a better example than rain.

DR. LINDLEY: Well, it is the probability of the temperature tomorrow.

DR. SOYER: And you can always define warm. I can ask you what you mean by warm. So you can give me a temperature. Then I can always redefine the event, then use probability theory.

DR. YAGER: But how do you define warm?

DR. SOYER: I can ask you what do you mean by warm.

DR. YAGER: What do you mean by warm?

DR. SOYER: But then I will ask your subjective opinion about it. For example, my notion of warm might be different from yours. I might say that warm is from 10 C to 15 C.

DR. YAGER: 9.98 is not warm?

DR. SOYER: It depends on your subjective opinion about what is warm.

DR. ZADEH: That is precisely the point. That is what fuzzy logic tends to get away from, the imposition of those artificial thresholds. It is not that 9.999 is cold and 10 is warm. That is the point. There is no sharp break. It is a matter of graduality.

So there is a degree to which it will be a warm day. So the truth value is a value between zero and one.

That is where this example that Professor Lindley was talking about would not work, because you can no longer say that it is true or false. If you want to, you can use multivalued logic for the assessment of the truth value of the statement that the event has taken place. That is the point.

DR. DeGROOT: Could you say a little bit about how the concept of learning would enter into your theory of fuzzy logic?

DR. ZADEH: Learning, of course, is a very complex concept in itself, and there is no universally agreed upon definition of what constitutes learning. But one simplified perception of what constitutes learning, which has been implemented to some extent, is one done by Professor Sugeno in Japan. He has done very interesting work where you have a rectangular track and you have a model car that can be steered by a human.

You steer it and it does certain things. Then a system takes over. The system which takes over has learned the algorithm that the person uses in maneuvering this car through this track.

The system does it automatically. No matter where you put the car, the system does it. So the system has learned how the human operator does that sort of a thing. This is something that has been done already.

The system, for example, may learn how to park a car. That is a little more complicated, the same sort of a thing. But the rules learned in that case are all fuzzy rules.

It would be impossible to do that sort of a thing using crisply defined rules. It's too complicated. This is part of what is called fuzzy logic control. Eventually the control is nonfuzzy but on the level of description of the rule, on that level it is fuzzy.

So you take a complicated situation and you try to describe the relationship between variables in fuzzy terms. That then is translated into fuzzy logic rules. Once it is translated then, it is implemented deterministically. That is the way it is done.

DR. SHAFER: I think this level of description idea, is a good point. But why couldn't that higher level of description be worked out in terms of probability?

DR. ZADEH: Too complicated. Let me give you an example that doesn't involve probabilities. It involves the description of a curve, for example. I want you to describe that curve.

I want you to describe that curve. If you look at a curve, qualitatively you can say when X is small, Y is large. When this is this, that is that. And in those fuzzy terms you can describe roughly this curve. There is no probability involved in that sort of thing.

Now you could describe this curve point by point but it is too complicated. You can capture the qualitative behavior of that curve by giving these fuzzy pairs: if X is this, then Y is that; if X is that, then something-something, and so forth which may be good enough for your purposes.

In other other words, that definition or characterization of a relationship between X and Y might be sufficient for purposes of control. But it is the fact that we are using imprecise characterizations that makes it possible, that makes it feasible to use a relatively small number of rules.

You see, if I ask you to define how you park your car, you will give me just a few rules. Those rules will be fuzzy rules. If I ask you to define it precisely, it will be an impossible problem. Too many rules would be required for that purpose.

DR. SINGPURWALLA: Is it true that when one subscribes to fuzzy logic one essentially abides with the calculus of probability, but finds it very difficult to work with, and therefore as an approximation one makes compromises and moves somewhere else?

DR. ZADEH: What you do is this: you certainly accept probability theory. You don't challenge probability theory in any respect. You merely say that the language of probability theory is too restrictive to deal with the imprecision that one finds in the real world.

So when Professor Lindley says that the probability that Shakespeare did not write Hamlet is .2, that .2 has a certain ring of precision to it that is not really justified. In other words, nobody can say that it is .2 or .3 or something.

In Rasmussen's report you arrive at the conclusion that the probability of a nuclear accident or something-something is something of 10 to the minus something. There is absolutely no way of verifying or proving or disproving, whatever.

In other words, it says that all of these statements are unrealistically and unjustifiably precise; that the most that you could say based on the information that you have is something like the probability that Shakespeare did not write Hamlet is quite low. That is the most that we could say really.

Any other numerical value is misleading. It conveys the impression of much greater degree of knowledge and understanding than you really have.

DR. LINDLEY: But you use numbers. In that example you used .3. Well, why is your .3 different from my .3?

DR. ZADEH: These numbers are on a different level. Probability uses a number but nevertheless it gives you a less precise characterization of a deterministic situation.

DR. LINDLEY: But you say my .3 could be .4. Why can't your .3 be .4?

DR. ZADEH: No, I wouldn't say .2. I would say low.

DR. LINDLEY: But you don't. You put numbers.

DR. ZADEH: In the definition of low.

DR. LINDLEY: That paper of yours has duodenal ulcer of .3. Now what I am saying to you is that you are committing yourself to numbers. If you are committing yourself to numbers, why is your commitment to those sorts of numbers better than my commitment to another sort of number?

They are both fuzzy. They are both crisp. I don't see the difference. The point is that your numbers combine in a peculiar way and mine combine in another peculiar way, but at least I can justify my peculiarity.

(Laughter)

DR. YAGER: In a certain sense, the difference is in the quality. When you say probability of .3, you are committed to one number, okay? But when you talk about a fuzzy set and you assign numbered values to it, you give a whole bunch of numbers which in a sense sort of nullify each other. Each number in itself is not as significant as the whole

bunch. So you could be off on one number and this doesn't have to be as precise as that one number that you give.

When you give .3, that is a very precise one piece of information. When you give a fuzzy set, you are giving a whole bunch of numbers. It is the total of the numbers that count and so if an individual number is off it really doesn't matter.

It is sort of the same thing as when in probability when you have a whole bunch of readings and you give the mean, you lose a lot of information than you would if you had all the values.

DR. LINDLEY: When you give a membership function for fuzzy sets, it is a probability distribution, which is a whole set of numbers. Exactly the parallel is between a fuzzy membership function and a probability function. They are both a lot of numbers.

DR. WATSON: I don't think that's right, Dennis.

(Laughter)

It seems to me that the situation you have to compare is your saying .2 and Lotfi's saying very small or quite small. He might articulate that quite small by a fuzzy set which will be a set of whole numbers. And it is his set of numbers which is being compared with your one number. What Ron is saying is that you can afford to be out quite a bit on one or more of this whole set of numbers. You can afford for your function to be playing around a bit and you hopefully won't have much in the way of a different conclusion.

But if your .2 were out and was .25 instead, it may affect the answer. But that of course is the crucial test for the difference between the two. What is the implication of the different theories?

It seems to me that what we need to do, particularly in fuzzy set theory, is to test what the output implications--how sensitive they are--are to these input membership functions.

I suspect that they should be quite sensitive and this does worry me. But I don't think you are right in the two things you are comparing.

DR. SHAFER: In the case of the fuzzy control story, if I got the story right, the numbers are actually pretty precise and they are gained from the experience of the calibrating trials. Is that right?

DR. ZADEH: No, that's the point. You use fuzzy control in situations in which you have a great deal of robustness. So coarse control is adequate. You wouldn't use that kind of control in a situation in which a high level of precision is expected.

So you essentially take advantage of the tolerance for imprecision. Let me give you a quick example of that sort of thing. Suppose you want to park your car. Now when you go to park your car, the final position of the car is not specified very precisely.

We want it within a few inches of the curb and the angle could be something-something. And humans can take advantage of this imprecision. So it takes very little time to park your car.

But if I specified the final position precisely, if I said the car should be within one-hundredth inch of something and it should be somewhere plus/minus three seconds of an arc of something-something, it would take you five years to park your car.

That is the point. So there are many, many situations in which there is this tolerance for imprecision. In whatever you do, you take advantage of that. If I want to go through this door, it is not that important as to whether I pass five inches on this side or five inches of that side, and my actions are governed and influenced by this lack of need for precision.

So this is essentially what you try to take advantage of when you use fuzzy logic control. It's the tolerance for imprecision.

DR. LINDLEY: But that is exactly captured by utility functions in the distance from the curb.

DR. DeGROOT: My understanding from what you are saying is I think we all agree. There are many situations where knowing that usually the situation is thus and so is enough to allow us to act. And no probabilist or Bayesians--or whatever we are being called these days--would disagree with that. I recognize, even though I am geared up to do a Bayesian analysis in a given problem, that I may not have to specify my probability distribution down to the probability of the last possible event, because as Dennis says, I know the utility function and I know the specific problem at hand would only require a few crude measures of probability.

And in those cases I am sure we would both do the same things. You say usually it is thus and so and so you are going to do it a certain way, and I say well the probability is pretty large on this and I don't have to do it.

I am certainly not going to waste my time doing calculations, as was suggested earlier, on a thousand dimensional parameter space when all I need is a very crude probability of a certain event to determine my action.

So I think that what you are saying is often interpretable in terms of probability. You are saying there is no need to gather very precise information in many circumstances, and I agree.

But what do you do when you do need more precision to choose a reasonable action?

DR. ZADEH: Use the classical probability theory.

DR. DeGROOT: Good. I didn't like the word "classical" but -

(Laughter)

DR. SINGPURWALLA: You mentioned the Rasmussen report and you mentioned the probability of accident or whatever it is that they were looking into with a certain number.

I think for the record I should also say that that particular number in the Rasmussen report had an uncertainty statement connected with it. It was done using a fault tree where they had uncertainties for all basic events and it was the propagation of those uncertainties using the regular calculus of probability that was used to arrive at the top number plus an interval around it. So uncertainties can and have been assigned to probabilities.

DR. ZADEH: I think you will agree with me that whatever intervals associated with it were excessively precise in the relation to our understanding of the whole thing. It could be way off.

DR. SINGPURWALLA: Oh, I agree with you that there may have been strains of optimism there, and I agree that there may be strains of handwaving and all kinds of other things that went in the Rasmussen Report. But the point is that it used the calculus of probability to assess uncertainty.

What is at issue here is the implementation rather than the philosophy of the logic which went into coming up with these numbers. I think what we are discussing is the means rather, than how it was done.

I agree with you that it may have been done in a sloppy way.

DR. DeGROOT: Well, to follow up on that comment and come back to the point that was made before, I think that if Dennis tells us that his probability is .2 that Shakespeare wrote Hamlet, and we say that that is a totally precise statement, no denying that, I think we are entitled to ask him, as some of us were asking before, how did he arrive at that .2. And I think we are entitled to know that if he is thinking of ten different possible contingencies, what probabilities would he assign to Shakespeare having written Hamlet under each of those ten possible contingencies. I think we are then entitled to know those ten conditional probabilities, as well as the ten probabilities that he is using as weights over which he is averaging them.

His .2 that he gives us as his final overall marginal probability is a weighted average of many probabilities. And so we want to know what those many probabilities are and what the weights are, and I think we are entitled to ask him for those. And we are entitled to disagree with whatever aspects of those 2N numbers are that he has to tell us.

That will permit us to think about how we want to think about the probability that Shakespeare wrote Hamlet. Do we want to simply accept his .2? Do we want to raise it or lower it because we disagree with some of the components that went into it?

I don't think to do a Bayesian analysis means that you only report a single number at the end and everyone operates from that. Not at all. Indeed it is your responsibility as a scientist to report all the probabilities that went into this final overall expectation.

Every probability is an expectation or a weighted average. So I think that should be part of it. That doesn't contradict the Bayesian approach to say that there are many numbers which enter into it.

Yes? You have been waiting very patiently.

DR. YAGER: Stephen brought up the whole idea of how you get these membership grades and things like this. It seems to me that the diagram that Glenn drew in his talk was very interesting about where all this stuff fits into expert systems in that we have a sort of natural language, and we convert that to some sort of mathematics. I think that is what we are really all doing here. What we're doing is manipulating the mathematics and then coming out with some sort of linguistic information at the end, and then it goes to some sort of user.

It just dawned on me it would be interesting to look at the fact that if you give a person who has to make some decisions--let's say somebody in the Navy, for example--if you give them information that says that the probability of the enemy doing this is .8 or if you give him the information that the probability of the enemy doing this is high, I wonder if he may be more able to deal with the fact that it is high than the fact that it is .8.

Somehow .8 is a very, very sort of lonely number standing out there in the middle of nowhere. Somehow I have the feeling that high or some linguistic information sort of tunes in much better to his own decision-making system. We have to remember to provide this information for users.

DR. FISHBECK: That depends on how he is trained. I submit that .8 can be very meaningful to somebody rather than high, slurring the Navy like that I guess.

DR. YAGER: No, no. I am saying any human being.

DR. GROSS: I'm Don Gross at George Washington University. I wonder, when Dennis came out with the statement that the probability that Shakespeare wrote Hamlet was .2, there was sort of an "uh."

Would it have made any difference to us if he had said the probability that Shakespeare wrote Hamlet was .25 or .15? I think the impact of the statement was that it was a low probability.

It seems almost that it depends on the application -- I don't want to say implementation--of the situation of which the statement is made. Would it have made any difference to us if he would have said .25 instead of .2?

(Pause)

DR. DeGROOT: Is there anyone here to whom it would have made a difference?

(Laughter)

Except Dennis.

(Laughter)

DR. SHAFER: I think the swine flu story is a good story which has been cited. It is very justifiably a story where numbers would have helped a lot because you had a communication problem where stories--it turns out that words do convey pictures to people and sometimes not the pictures that were meant to be conveyed, and it goes from one person to another and it gets changed more easily if it is words than if it is numbers.

DR. DeGROOT: Mr. Wise?

DR. WISE: I asked this question originally when Professor Lindley just started talking and I can't resist bringing in a little point from physics. That is, in quantum mechanics it is considered a big discovery that electrons do not act like little painted balls and urns.

And you can't characterize them with single numbers. At the minimum, you need complex numbers. And that is a formalism they developed to try to explain their experiments. They did an experiment to prove that.

And if we are worrying about philosophical foundations, we have yet to demonstrate that one number is sufficient. Maybe it is a triple. Maybe it is a triple of complex numbers. In some case, one number won't do.

DR. DeGROOT: Dennis?

DR. LINDLEY: The situation there as I understand it is as follows. Suppose that you did give a pair of numbers instead of one number and suppose that you scored that pair of numbers.

Then it follows that you are redundant and really you need only have given one: the probability.

But suppose you give two numbers and you use two score functions at the same time, okay? What these two score functions are trying to do is to express somehow different qualities of the system.

If you used two score functions--I must say I haven't followed through the mathematics--but it looks pretty clear, of course, that you would, in fact, finish up with two numbers. What their rules would be, I don't know.

But a single measure of worth leads to probability. Two measures of worth will lead to something more complex.

DR. WISE: No. I would argue otherwise because in quantum mechanics you have a complex number and the magnitude of the number is in fact probability. And so your argument applies very neatly to the magnitude.

But in order to work with them and add them, you have to have the interference reinforcement effects you get with complex numbers. You can't get the correct probabilities at the end unless you do all the calculations in the middle with complex numbers even though you are scoring probabilities at the end.

DR. LINDLEY: Yes. I am afraid I can't argue about the technical--

DR. WISE: My question is why do you think that is a unique situation? Why do you think in all other cases a single number is sufficient?

DR. LINDLEY: Well, I don't think a single number is necessarily sufficient. What I am saying is that if you are prepared to do it by means of a single number, and as I see it, despite the statements which have been made, belief functions and fuzzy people do use a single number. That if you use the single numbers, those single numbers must be probabilities.

Now if you are going to use pairs of numbers or more complicated things and you only use one score function, then I am sure it comes back to probabilities. If you use two score functions, then I am not clear what happens but yes, you may need two numbers.

DR. SOLAND: I can inject something about expert systems which I don't know very much about. I do know from the paper of Professor Zadeh's that I read that there appear to be some applications of fuzzy set theory which sound very much to me like expert systems in terms of control.

So there is some proof in the pudding that it has worked, at least for expert systems. Do we find the same thing with a complete Bayesian probability analysis and belief function analysis? Can we point to some working expert systems or prototype expert systems based upon these? And what has to be done to make them better or to get them if we don't have them yet? Professor Shafer didn't give enough detail

for me to follow about what the difficulties were in using them, probabilities for example, in some of the expert systems that he talked about. But if there are operational difficulties, then I think we ought to discuss those because maybe one of the big benefits of the fuzzy set approach is that it avoids certain operational difficulties.

DR. DeGROOT: That's a good point. The subject of the conference is, at least in part, expert systems and that would be interesting to hear some more about the operational relationship of these.

DR. ZADEH: Of course, many of the people who work in fuzzy logic, myself included, were brought up on probability theory. Most of my best friends are probabilists.

(Laughter)

So it is not a strange subject. I include Professor Lindley in that class. So it is not something that you are not aware of. The point is that contrary to what is accepted in the case of the classical probability theory, you do make a differentiation between something that is imprecise in the sense of being possibilistic, and something that is probabilistic.

In that case, "high" interpreted as a possibility distribution is a generalization of an interval. An interval is not a probability distribution. By "high" you mean more than so many degrees, or so many inches, or whatever.

When you take high to mean something like possibility distribution defined by one of these curves, that is an extension of an interval. It is not an extension of constant probability.

It is precisely because of the lack of differentiation between the two--possibility and probability--that we find ourselves in situations where, contrary to what Professor Lindley said, there are many problems that cannot be handled within the framework of probability theory.

The rules of combination are quite different. You don't combine possibilities the way you combine probabilities. And there are many, many examples of situations in which if you interpret these things as probabilities you get completely wrong results or else horribly complicated results. It is one or the other.

So now the point that was made here by David is that it is not the matter of acceptability. Of course everybody would prefer to have a crisp number of .8 to high. It is a question of justifiability.

It is really justified to say .8 based on the information that you have. I would like to return to the point you made and I think it is a very good one--I hope Professor Lindley will not take offense to that.

I am pretty sure if he stood in front of a blackboard and explained to us the way that figure of .2 was arrived at, we would not accept that kind of an analysis.

We would see that it is shot through with all sorts of assumptions and this and that. And in the end we would say look, .2 has no justification whatsoever; that the most you could have said in terms of intervals is between zero and .5, or in fuzzy terms that it is low. That is the most you can say.

Most of us also would like .2. So that what we have to differentiate is between acceptability and justifiability. It is a very different thing. Is it justified to be that precise?

In the case of classical expert systems, MYCIN and PROSPECTOR, my contention is that whatever answer they come up with, those certainty factors are unjustifiably precise. Unjustifiably precise. But people like that.

DR. SPIEGELHALTER: I don't want to jump to what I am going to say tomorrow, but for any point probability that anyone gives out, one should definitely give ranges, and there are at least three different types of ranges that one could give around it.

Certainly a point probability on its own is certainly without any justification for it. But there is no need to deviate from probability theory in order to provide some idea of the possible variation around that point probability.

If one has to act, then the point probability is the one that one should use. But in order to justify it to someone, then it is quite reasonable that the possible variation in that probability by a slight change in the analysis, by the imprecision of the inputs, can be given out as part of the output.

That again is like it says high, and you can say what does high mean. Similarly in a probabilistic system, it will say .2 and you say what does that .2 mean? And it will give you a distribution around that .2 and tell you what the distribution means.

DR. ZADEH: Is it a probability distribution?

DR. SPIEGELHALTER: Well, there are different distributions one could give. One can give a probability distribution.

DR. ZADEH: Then you are talking about second order probability?

DR. SPIEGELHALTER: Yes, essentially. What I am saying is there is the hierarchy of uncertainties about the imprecision on the probabilities, and whether or not it's represented by the second order probability distributions or whether it's represented by fuzzy calculus is an imprecise number that's put on there.

If one is really looking for the meaning, the only things I understand are probability distributions and the only thing the people I work with understand are probability distributions.

I would not want to use a system that generated something that I could not explain. One thing that has been mentioned, in Dennis's talk, is the idea of external calibration of probabilities. Can they have a meaning calibratable against events in the world?

Okay, everyone has to bring in some idea of a long-run frequency to that argument, which is perhaps unacceptable to really pure Bayesians, but there is that idea of meaning that can be given to the numerical outputs. And I find those concepts of meaning and justification missing in linguistic output from an expert system.

DR. YAGER: You say the only thing you understand is probability distributions. In point of fact, possibility distributions are all over the place. A perfect example of that is think about linear programming, for example.

Are you familiar with linear programming? You have linear programming and you have some function you want to maximize, and you have some constraints. You cut off this space and you say what solution optimizes this.

And before you do the operations you say well it has to be something within this space of possible solutions. Some definitely can't be and some definitely can be.

That is a possibility distribution, albeit one that just says zero/one membership grade, but that is a perfect example of possibility distribution. Then if you look at the objective function maybe you could sort of say the answer can't be over here, it could be over here, and may be over here; and maybe you could get some other numbers other than zero or one. But that is a possibility distribution.

DR. SPIEGELHALTER: Sounds like a restriction of a range. One can state one's own personal uncertainty as to where in that range the thing is.

DR. SINGPURWALLA: You're uncertain where the solution lies. You can give a probability for the solution lying on each corner before you solve the problem, so I don't see any difficulty with linear programming.

DR. DeGROOT: Let me try. I think that he's not talking about second order probabilities. What we are talking about--and I agree with much of what Lotfi Zadeh said--but I don't like the term that I wouldn't agree with Dennis's probability of .2 because it is unnecessarily precise.

I may not agree with that because when he told me his argument I would see there were many other possible probabilities and if I wanted to think about it I would try and determine some for myself and might arrive at a final overall probability quite different from .2.

I think what that means is that the .2--not that it is unnecessarily precise--just because I wouldn't agree with it, but that it is very sensitive. There are some probabilities that are insensitive, insensitive to learning, to further data.

If Dennis says that he has studied this for ten years or more and studied the problem and looked into all the possible relevant sources of information and so on and he knows whatever there is to know about the subject, the probability is .2 and there is very little that anyone else could tell him at this point that would change those probabilities, that is one kind of probability .2.

There are many other kinds of probability .2's, namely very sensitive ones where any little bit of further information could change that probability dramatically. And I believe that is what you would refer to as a situation where you really need some sort of a fuzzy statement. I wouldn't say the probability is unnecessarily precise. I would say it is very transient in nature and very sensitive, in a sense, to any other little bit of information. If I go home tonight and I think about it, it would change from .2 just by remembering Shakespeare from high school and God knows what.

So maybe there is not much point in specifying an exact number if it is going to change very soon anyway. I mean we all know the example about the probability is .5 of getting a red ball from the box because we have no idea of the contents, and the probability is .5 about getting a red ball from the box because we are certain that exactly half the balls are red.

They are both .5. To me they mean exactly the same thing but I certainly recognize that one is very sensitive to further information and one is totally insensitive to further information.

I don't think you have to get into second order probabilities, but just if we recognize that our overall final marginal probability is a weighted average of some things.

In one case, it is a weighted average of wide variety and in the other case it is a weighted average of a very tight, tight range. Maybe we will settle these issues because maybe we are all trying to say the same thing in somewhat different languages.

DR. DeGROOT: David, don't give your talk for tomorrow now.

DR. SPIEGELHALTER: Okay. I will be expanding on exactly that tomorrow.

DR. DeGROOT: Oh, I'm sorry. I didn't mean to give your talk for tomorrow here.

(Laughter)

I think we have time for just a couple more questions.

DR. KONG: I think I agree totally with what Professor DeGroot has said. I have been looking at the belief functions for a few years and I have been like using sort of something like upper and lower probabilities and I never pretend to say that because I have two numbers it is better than one--of course something like an interval of .1 and .4 cannot be more precise than something like .3. In the sense of what is the precision of the value .1 and .4? In fact, sometimes I may prefer something like an interval like the evidence that the belief function is based on--I think this is something which Glenn has written on for a long time.

Let's say if I have a belief function which has .1 probability supporting the proposition of A and .4 probability supporting the proposition not A, then basically I've got an interval for the proposition A, something like from .1 lower probability to 2.6 upper probability.

So I have a range. But another way to look at it is I actually start with a Bayesian distribution function which has like .2 and .8 probability, which is a Bayesian distribution because it adds up to one.

But then I reevaluate my evidence and sometimes it seems that my evidence may not be relevant in this situation, and I say maybe about .5 of the times I think this piece of evidence may not be dependable at all.

And when this piece of evidence is not dependable, then I don't know anything at all. Then what we would do is basically we would discount the Bayesian distribution of .2 and .8 by sort of a factor of .5.

Again this value .5 is not really precise. It is not exactly .5. I may mean something around .48 but I just picked .5. So by doing that I end up with what I originally have. I have the range from .1 to .6.

And basically what this will do is when I have other evidence, when I come up with other evidence which is sort of maybe conflicting, sort of a point towards another direction, then if I have a discounting factor which is big, then it will be much more sensitive to the new evidence. It will be dominated by the new evidence, especially when it is very contradicting.

DR. DeGROOT: Well, we will talk about that tomorrow, discounting factors. I have some questions about those, too.

DR. KONG: So basically I just want to say that using two numbers doesn't imply that we think it is more precise than one number. It is not that at all.

DR. DeGROOT: Mr. Wise, a short comment, please.

DR. WISE: Real quick comment. If you are using upper and lower probabilities and you are trying to make decisions and look at basic decision theory, it can make a big difference whether you assume that the probabilities are uniform, in-between, or skewed to one end.

And if you just strictly use the upper and lower probabilities, you don't know and it could greatly affect your decision.

DR. SOYER: Whatever the value P is, it is just his subjective probability. The argument on precision then is a problem of the decision-maker when he is evaluating Professor Lindley's probability. So according to the decision-maker's belief about his expertness, he can always change that probability by making a conditional statement. So it is just a matter of evaluation of the probability forecaster or predictor.

DR. DeGROOT: Okay. Well, I think this is a good start for tomorrow's discussions. I do want to say something important before we close. That is, it has been a long time since I have stayed awake for a whole day of talks and I did today, and I think it was not only because I had to be awake at the end to stand up here and do this, but also because all three talks were really excellent.

They were pitched at a level that I could understand. They were clear and I was very impressed and I learned a lot about presenting good talks today and I do want to congratulate--I hope you will join me in congratulating all three speakers. Thank you.

(Applause)

GENERAL DISCUSSION II*

DR. DeGROOT: Are there other questions pertaining to David's talk? We welcome general comments also at this time. I would like to reserve the hour after lunch to give each of our four speakers 15 minutes or so to give their responses and general impressions of these two days. If there are other comments prior to lunch, I would welcome hearing them. In particular, I again renew my invitation of those of you who have comments about the practical implementability of the various systems. I know we would like to hear those. Art, do you have a comment?

DR. DEMPSTER: I have had lots of turns, if there are others. I could raise a topic.

DR. DeGROOT: You win.

DR. DEMPSTER: In a sense I will follow up a little on Morrie's issue about the variability or refinement of probabilities and will try to relate it to something that Lotfi Zadeh has been saying. Lotfi has raised the issue a number of times, about complexities somehow are an enemy of probability or the believability of probability.

And there is a sense in which I am sympathetic with that idea. I feel that as an applied Bayesian statistician when I make the models more and more complex, I perhaps trust them less as I have to do, and he states somehow as though it is a consequence of fuzzy logic that somehow the probabilities will become more confused. They will just have a range zero to one and will be essentially useful and I am interested in knowing, and perhaps he can elaborate on this sometime, in what sense is that defensible? What sort of logic is used to draw that conclusion?

I think the issue is important in part because in Bayesian theory, if you have a Bayesian model, the more information you get about the patient in some sense, the better off you have to be. You have a more refined judgment and I am sure mathematical criteria measures of information could be created that showed that in fact you get more information. So in that theory more information is always better, but apparently according to fuzzy logic, more information can be worse.

So there is a kind of paradox here which might help to resolve the difference between the approaches. I think again that belief functions are very likely on the same side as probability in this dispute, although I am not familiar with the technical aspect of belief functions which would argue that more data is more information in the probabilistic sense.

---

*Session followed Dr. Speigelhalter's presentation and its discussion

So I just wanted to raise this issue for further discussion.

DR. DeGROOT: Gabe.

DR. PEI: I don't know if some of your comments were directed toward me, but I have an example of an application of Bayesian modeling. It is an area I've been working in for the last several years, having to do with the Navy, particularly with the search for moving targets. The approach has been extremely basic, where we come out with prior distributions for the target's location and model detection capabilities as carefully as we can, and then we update posterior distributions for the target.

On paper it seems to hold together very well and the individual parts of the model seem to calibrate very well. We can, for example, calibrate the capabilities of sensors. We can, for example, validate the behavior of moving targets. When you want to assemble all of these models together and try to do prediction or try to optimize various plans, we have difficulty and probably some of the problem is due to the stochastic nature of the problem, the thing that evolves over time. We have tried to come out with stochastic models. What you do is try to take that into account. But they never seem to behave in exactly the way we expect them to behave when we go out and try to use them.

Part of the difficulty I think has to do with the amount of information that we think is out there and we have come up with plans. We have come out with predictions which after a period of time tend to be very precise. As a matter of fact, even if you get no observable detection, that is still information in the sense of where you think the target is not. And so after a period of time your estimates become very, very precise and yet many of the times they are completely off, totally off. That is simply because over any reasonable stretch of time the models tend to break down.

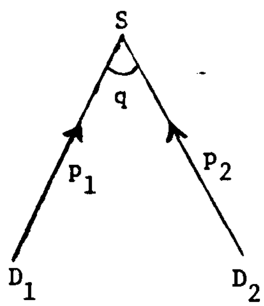
Now, we can try to patch the models together by adding more parameters. But it turns out that human operators do a lot better in these models if they can interpret the process at any time to re-initialize the problem at any time and do a lot better than any automated process which we can build into that.

So that is a puzzling one -- well, I don't know if it is puzzling, but it is an aspect of trying to use probabilistic statistical methods to a very practical problem which tends to have certain limitations. I don't know what the solution to this is. I don't know if anyone has any comments on that.

DR. DeGROOT: Thank you. Mr. Kong, do you want to describe some of your work? This might be a good time if you would like to do that, maybe take about ten minutes or so.

MR. KONG: What I am working on is a causal medical diagnostic model. The basic building block of our model consists of a symptom S and diseases D_1, D_2, ..., D_n, which are possible causes of S.

Consider the simple case where there are only two possible causes, D_1 and D_2. The relationships between the diseases is represented by the following diagram



(Diagram 1)

The arrows pointing from D_1 and D_2 to S indicate the causal relationships. Probabilities p_1 and p_2 are attached to the arrows. There is an arc between the two arrows with a probability q attached to it.

The number p_1 is interpreted as a causal probability. It is not the conditional probability of the symptom given the disease D_1. It is the probability that D_1 causes S given the presence of D_1. Thus it can be thought of as the lower probability of S conditioned on the presence of D_1. This is because if D_1 does not cause S, S can still be present because of other causes. What we are doing is attaching a probability to a logical statement. Logical statements can be represented by sets as illustrated in Table 1. (For notations, let $\mathcal{D}_1 = \{d_1, \bar{d}_1\}$, $\mathcal{D}_2 = \{d_2, \bar{d}_2\}$ and $\mathcal{S} = \{s, \bar{s}\}$ be the outcome space of D_1, D_2 and S respectively, with d_1, d_2 and s denoting the presence of the diseases and the symptom, and $\bar{d}_1$, $\bar{d}_2$ and $\bar{s}$ denoting the absence of the diseases and the symptom.)

Table 1

Probability attached	Logic	Set Representation
p_1	$d_1 \rightarrow s$	$\{ (d_1,s), (\bar{d}_1,s), (\bar{d}_1,\bar{s}) \}$
p_2	$d_2 \rightarrow s$	$\{ (d_2,s), (\bar{d}_2,s), (\bar{d}_2,\bar{s}) \}$

Consider the logical statement "d_1 implies s". If d_1 is the outcome, then s must be the outcome. On the other hand, if $\bar{d}_1$ is the outcome, then both s and $\bar{s}$ are possible. This explains the set which corresponds to "$d_1 \rightarrow s$" in Table 1.

DR. WISE: Are those last two pairs (referring to the first line of Table 1) supposed to be identical? Is there a bar?

MR. KONG: No, they are not identical. There is a bar here (referring to the bar over s in the last pair). The product space corresponding to D_1 and S has four elements. Here I allow three of the four. The pair $(d_1,\bar{s})$ is ruled out. The situation is similar for d_2; d_2 implies s with probability p_2. The only difference between line 2 and line 1 of the table is that all the 1's are changed to 2's.

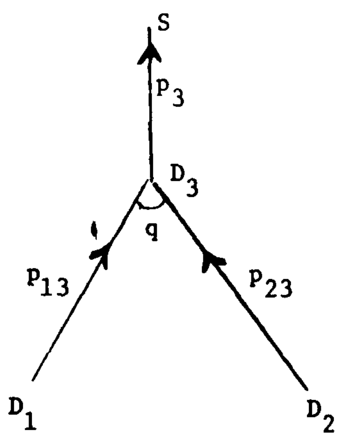
Now we come to the arc and the number q. The arc stands for the logical statement "s implies at least one of d_1 and d_2" with probability q attached to it. Both diseases can be present. Notice that the logical statement "s implies at least one of d_1 and d_2" is equivalent to the logical statement "d_1 and $\bar{d}_2$ implies $\bar{s}$". The set which corresponds to this logical statement is the product space $\mathcal{D}_1 \times \mathcal{D}_2 \times \mathcal{S}$ minus the element $(\bar{d}_1, \bar{d}_2, s)$.

The arc and the arrows each corresponds to a belief function. For example, the arrow pointing from D_1 to S represents a a belief function with two focal elements. They are $\{ (d_1,s), (\bar{d}_1,s), (\bar{d}_1,\bar{s}) \}$, the set in Table 1, and the product space $\mathcal{D}_1 \times \mathcal{S}$. The basic probability assignments are p_1 and $1 - p_1$ respectively. Assuming that the p's and q are independent probabilities, the belief functions corresponding to the arrows and the arc can be combined over the product space $\mathcal{D}_1 \times \mathcal{D}_2 \times \mathcal{S}$ using Dempster's Rule. Renormalization is not necessary when combining these belief functions because there is no conflict between them. Also notice that the belief functions are all purely relational, meaning that the marginal belief functions of D_1, D_2 and S are all vacuous.

DR. SINGPURWALLA: When you say these probabilities are independent, what do you mean?

MR. KONG: It means that the diseases are considered as independent causes of the symptom. Consider this (referring to diagram) as an example. The diseases D_1 and D_2 are independent causes of S. Each of D_1 and D_2 can cause S, but they do not interact with each other. For instance, if both D_1 and D_2 are present, then the probability that they will not cause the symptom is $(1-p_1)(1-p_2)$.

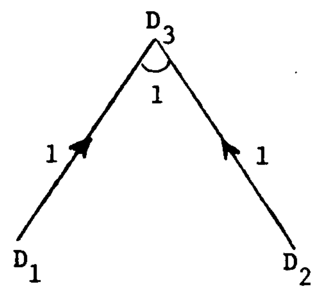
Dependent causes can possibly be modeled by something like this



(Diagram 2)

The node D_3 can either be real, meaning that there is actually a disease D_3, or artificially constructed. The numbers p_3, p_{13} and p_{23} are probabilities attached to the corresponding arrows in the diagram. Here D_1 and D_2 are dependent causes of S because they are both causes of D_3 which in turn is a cause of the symptom S.

A special case of this is the problem of disease class, which I think both Glenn and David have talked about before. Assume that diseases D_1 and D_2 form a disease class which we call D_3. The relationships between D_1, D_2 and D_3 can then be represented by this

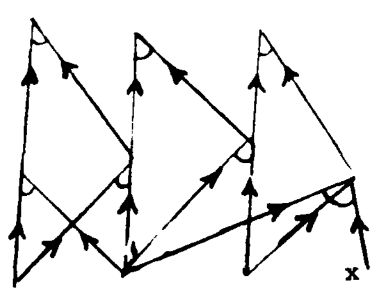


(Diagram 3)

The p's are both 1's because if the patient has one of D_1 and D_2, he must have the disease class D_3. Similarly, q is 1 because if the patient has the disease class D_3, then he must have at least one of D_1 and D_2.

In the case of disease class, sometimes we may want to simplify the problem by assuming that the patient has one and only one disease. In our model, this is not necessary. We make this additional assumption only if it is reasonable. In most situations multiple diseases should be allowed. Our model is quite flexible in this respect.

These causal structures (referring to diagram 1) are just building blocks of a model. The models I have been studying are called layered-models. The diseases and symptoms are grouped into layers, something similar to what David has talked about. The idea is that we only allow arrows pointing from a node to nodes which are located one level above it. So the model may look something like this



(Diagram 4)

This model is more general than the tall graph that Glenn has described. If we ignore the direction of the arrows, there can be loops in the graph. On the other hand, if we follow the directions of the arrows, we won't run into loops.

I arrange the nodes into layers mainly for the purposes of implementation and computation. The general case will be that we have a set of nodes and a set of arrows pointing from nodes to nodes. If it is true that there are no loops if we follow the directions of the arrows, then we can always rearrange the nodes (maybe some artificial nodes have to be added) such that they form layers.

The arrows and arcs in a graph correspond to simple belief functions over the product space. Theoretically, combining these belief functions over the product space will give us the global relationships of the diseases and the symptoms. On the other hand, the amount of computations required may make this practically impossible.

What we are interested in doing is the following. We have a patient who is observed to have certain characteristics, the absence and presence of some symptoms. Based on the observations, we want to find the conditional marginal belief functions of some diseases we are interested in. The major roadblock to this task is again the amount of computations required. Fortunately, the amount of computations can be reduced by taking advantage of the layered structure of the diseases and the symptoms.

Consider a two-layer model where the bottom layer consists of diseases and the top layer consists of symptoms. The diseases are marginally independent, meaning that if nothing is known about the symptoms, then the observation of one disease does not provide information about another disease. This property does not hold for the symptoms. One way of thinking about this is that information can propagate downwards and then upwards, but it cannot propagate upwards and then downwards. Because of the latter, the amount of computations can be reduced.

But even though we have this kind of structure, actual computations of the joint belief function will still be impossible in most cases. The approach I am thinking of right now is the Monte Carlo method. The Monte Carlo method may work in this case because we are interested in some marginal beliefs instead of the joint belief function itself.

If we have 100 diseases and symptoms, the product space, which is also the frame of discernment, has 2 to the 100 elements. The belief function over the product space is defined by 2 to the 2 to the 100 numbers. Since we are only interested in the marginal beliefs of a few diseases, we can use the Monte Carlo method to estimate only those numbers we need.

DR. DeGROOT: Thank you very much. No, we don't have time to run through the two to the two to the 100 right now, but after lunch perhaps.

(laughter)

DR. DeGROOT: Are there further comments or questions?

DR. ZADEH: There is a branch of logic called inductive logic, which is very closely related to this whole thing. With inductive logic you have your associated probabilities with formulas. For example, P_1 implies P_2 or P_1 and P_2, or whatever. We have formulas. So you associate the probabilities with each formula and then ask what will be the probability of some other formula.

Generally, you come up with bounds. That is the standard thing, which relates to Dr. Shafer's theory in that you associate probabilities not with atoms but with formulas, with propositions and that is why this resulted in what Augustine just said. Basically what happens in that logic, very closely related to that, is that you come up with bounds on the probability of a given proposition or formulas along those propositions and generally.

And this is the work that some people in AI have become interested in more recently. They call it probabilistic reasoning. Probabilistic reasoning is very closely related to inductive logic and some of the basic papers on this subject go back many years. Also, Barry Adams has written a large number of papers dealing with the question of propagation of probabilities from the premises to the final conclusion.

DR. DeGROOT: Are there any comments over here? Alright, let us adjourn for lunch. After lunch I will call on Glenn Shafer, Lotfi Zadeh again and Dennis Lindley, give them each about 15 minutes or so and David also, although we have heard from you more recently, and give them a chance to summarize their views and differences with the others.

Thank you.

CONCLUDING DISCUSSION

DR. DeGROOT: Thank you. This session is scheduled to give our speakers a chance to give their reactions to the proceedings and I would like to re-introduce Glenn Shafer.

DR. SHAFER: When Nozer was encouraging me yesterday to try to make a case for belief functions I sat down and tried to think of some things I should say and I am not sure I have produced the strongest case I can. I think it may be more of a case of taking the opportunity to say things the second time.

I thought I would start with a caricature of a 150-year-old controversay in statistics, the controversy between Bayesian and frequentist's points of view. One way to caricature that was the frequentist view wants to only look at the probability judgments that are based on observed frequencies, which is a very ideal particular kind of evidence. It is not only observed frequencies, but clearly relevant observed frequencies.

The Bayesian view is a little different, especially if you look at not the Bayesian view as it might have been, not if you look at what we now think of as Bayesian ideas, but if we look at the Bayesian philosophy: De Finneti, Savage and Ramsey; this philosophy seems to not pay any attention to the quality of the evidence at all.

One view says if we do not have this very special kind of evidence we can't make probability judgments at all. This other view is that the quality of evidence does not matter if we can write down events A and B, then write down P of B and then we should be able to make a judgment about that, whether we have good evidence for it or bad evidence for it. The philosophy does not have any place in it for discussing that.

Presumably that is left, in some sense, to the pragmatics on the subject. It is just not in the philosophy of the subject. I want to pose to you the question "how can we possibly find the middle ground between these two extremes?" It seems to me we need a vocabulary, a way of talking that naturally leads a little more to a middle ground. I think the way of talking that we need is to talk about constructive probability, to emphasize the fact that probability judgments are the things that we can construct based on evidence.

Another way of putting that is to emphasize that a probability analysis is only an argument. I have tried in some of the things I have written in the last few years to give some depth to that by talking about comparison to canonical examples. In my talk yesterday, I was talking about how belief functions have one set of canonical examples and Bayesian calculations have another set of canonical examples.

What you are doing when you make a probability analysis, it seems to me, is you are taking a natural situation and making a comparison to those canonical examples and you are usually doing that by comparing pieces and then trying to put things together to fit that picture. It seems to me that process does constitute an argument, because you are saying, look, knowing this is like knowing that there was a 20 percent chance of this happening and knowing this is like knowing there is a 30 percent chance of this happening if such and such is true.

The element of argument there is that maybe it is convincing to say it is like knowing that and maybe it is not convincing. Maybe your evidence does seem to have that strength and structure and maybe it does not. When Professor Lindley was talking about his 20 percent probability for Shakespeare writing Hamlet, what was it we wanted him to give us? Professor DeGroot gave one way of explaining why we were not happy with what he was saying. Professor DeGroot's way of explaining it was, well, we would like to see him break that down into a weighted average of conditional probabilities so that we could go and look at it and perhaps give some different values there and come up with our own judgments.

I think that is right as far as it goes, but I would like to go a little farther. I think what we really want to see there is an argument. We want to see what Dennis' evidence is and we want to see how he is putting it together and what kind of argument he has for his high degree of doubt that Shakespeare wrote Hamlet. It is not necessarily the case of having seen his argument I will feel that I can produce an argument for myself, because I may not be able to. I may not find the whole analysis, the whole set of supporting evidence. I may not find anything convincing about it at all. I may not be able to find it and put my own numbers in there.

But what I want to do is see that argument, to give myself a chance to be convinced. One aspect of this idea, that a probability analysis is only an argument, is just what I said, that there may not be any argument. It may not be the case that there is a right way of analyzing this evidence that is convincing and in probability terms we can produce either a Bayesian argument or belief function argument. Maybe there is not anything convincing or that is going to be convincing there.

So the general slogan that I would use to summarize those ideas is that probability is constructive. There was a remark made by, I think it was Ben Wise yesterday, and repeated by he and Terry Ireland today, in a discussion we had over dinner, which I have tried to resist and I think I should talk a little bit about.

I better get the slide. Okay, we are back to Fred. Fred is 80 percent reliable. Fred came up and told me the streets outside were icy. I think that Fred is 80 percent reliable. Eighty percent of the time when he wanders up to me and says something, he knows what he is talking about and he is telling me the truth. So that 80 percent is an argument which I would think of as an 80 percent reason for believing that the streets are icy outside.

Now, my attitude is that I think that is a good argument. It is only an 80 percent argument, but that is what it is worth and I am willing to talk about, since I am willing to talk about constructing probability judgments. I am willing to talk about thinking of that evidence alone and making a judgment on the basis of it, making an 80 percent judgment on the basis of it.

Now, you might say, what about your other evidence? Well, my attitude is I might have some other good evidence about this question and I might not. If we are willing to take this point of view that evidence does differ in quality, then we have to contemplate the possibility that the other evidence I have about whether it is icy outside may not measure up to the quality of Fred's testimony. It may be that I could make probability judgments about other items of evidence and bring them in and strengthen my argument. That may be the case.

On the other hand, it may be that the quality of the other evidence I could bring to bear is so poor that putting it together with my judgment about Fred's reliability would only weaken my argument. So that is the general point I make.

So the Bayesian analysis that I show you on the screen -- there are many Bayesian analyses, but the one I suggested on the screen, in my talk yesterday, had two additional judgments. One was a prior probability that it was icy outside and the other is the probability that Fred would be accurate if he is careless. I did not emphasize, historically it is usually number one for arguments about Bayesian and non-Bayesian methods, the prior probability argument. I think that is a little bit artificial.

Prior probability seems basically to refer to other evidence. It may well be the case that I have other evidence. In the second version of the story I told, I had a thermometer which I thought was better evidence than Fred. So it may well be that I have other evidence and that I should carry the argument further and take that other evidence into account. I think it is much more likely that that would be the case than that I could supply this number two here, the probability that Fred will be accurate if he is careless. I mean that is just personal, since I made up the story. You could make up a story you like better.

But it does seem to me that I can very well imagine having a general impression about a person's reliability and being willing to compare my situation to sort of a random draw from the different situations where he is reliable and not reliable. I can well imagine that being convincing to me, while it would not be convincing for me to try to model just what is going on in terms of what his chances are of accidentally hitting the truth, if he is being careless. That may just seem much more speculative to me and I may not feel like I can make a convincing comparison of that part of the situation to the picture of chance.

That is why I might feel that it would just weaken my argument to have to make a judgment there. These are the basic kinds of attitudes that I think you have, to have found an interest in belief functions, because it is the case that you can always make, given a belief function argument. You can always make a Bayesian argument that will take into account what you are doing, as a belief function argument, and it will also take more things into account, which will undeniably be a better argument if you could get all of the pieces to make it go.

So I think the only reason for being interested in the belief function argument is hearing that somehow these incomplete models as arguments may be of some interest. You just don't have the strength of evidence to make the more beautiful construction in the sense that there is any question that the Bayesian logic (if you have all of the pieces to put it together, just on that side) is much more; because of its greater completeness, it has much more convincing logic to it.

Now I come to Ben Wise's thought. Wise said, well, if we could see what Bayesian judgments are needed to get to the answer of eight-tenths, we will better understand the belief function analysis. I am going to resist that. Because I mean what would we need to get to the eight-tenths? I think what we would need would be prior probabilities of a half each for whether it was icy outside and the probability -- I was mentioning these judgments. To get a Bayesian analysis, I would need a prior probability and a probability that Fred would be accurate if he were being careless. I think the prior probabilities are half and half and if the probability he is being accurate if he is careless is zero, then I have the eight-tenths.

I just want to resist the idea that when I just take Fred's reliability as my reason for believing that it is icy outside that I am doing this. I don't think I am. I am not making these judgments. I am not making this more complete probability model. I am just depending on an argument that says that if something happens 80 percent of the time, and this is an example of that, that that represents a relatively strong argument for it. It is only an argument.

There is a feeling that if we do a Bayesian analysis we have more than an argument. When I say, yes, this just sounds like an argument; it does not sound like a complete analysis. But my point is that the Bayesian analysis always gives you an argument too. So what are my reasons for trying to resist this suggestion? Because the logic of the belief function and the Bayesian analysis are different, so I don't want to interpret them, the belief function analysis as a Bayesian analysis.

In trying to explain why the logic is different, I want to say things like the two arguments make different comparisons to pictures of chance and maybe I could convey that to you by saying that they imbed the problem in different sequences of problems. I mean the pull of probability always has this idea that you can imagine that this is -- you can always go from the subjective to a more objective picture where you really did not have a repetitive situation.

But the belief function argument in this simple case is obviously just looking at the repetitive situation of different things Fred says to me when he comes over to my desk with this funny look on his face. The Bayesian analysis is looking at an imaginary sequence of repetitions, is a much more precise story corresponding to what is happening right now.

So there is a different sort of sequence of problems in which this thing is being imbedded in and somehow the fact that the two analyses agree on a particular answer at this time should not be given that much weight.

I don't know, this is changing horses a little bit in midstream, because it is looking at a different example. Let me throw this out, since it seems to be an example that is easy to follow. I have used it in several papers. George Hooper was the Bishop of Bath and Wells in the 1680s. I will tell you a little about George Hooper really. There was a paper in the Transactions of the Royal Society published in 1699 on probability, the authorship of which was unknown to the probability community. Part of the history of probability called it anonymous papers and speculated that John Craig might have written it, et cetera, et cetera, and people repeated that for many years.

It turns out that all of the time the probabilists did not know who wrote this paper and the theologians knew perfectly well. In fact, they had republished the paper in Bishop Hooper's collected works, which were published in the 1700s and again in 1855. It is an interesting case of what is known in the sense of whether something is known or not. It was known to some people and it was not known to other people.

As far as I know Hooper is the first person to refer to a number between zero and one as a probability, to use that name for the numbers between zero and one. Now, that may not be right, but I can't name to you anyone who did it earlier and Hooper did it in 1685. He was a chaplain to King James II, I believe. Since King James II was a not so secret Catholic, being his Protestant chaplain carried political responsibilities, and one of the things he did in the course of that responsibility was provide a tract against the infallibility of the Pope and that is where he first published this argument that he was interested in.

So it was a question of combination of witnesses. And this is really the other point of this, that it proves that the belief function is older than Bayes, because Bayes did not write his essay until -- well, it was published posthumously in the 1760s and he was not born until the 1720s. So here we have these two witnesses who have their credibilities. P_1 is the probability that the first witness is going to be truthful as opposed to being careless. P_2 is the probability the second one is going to be truthful as opposed to careless. So, we have these two independent witnesses and they tell us they both agree on something they tell us. What probability does their concurrence give to the conclusion?

So Dempster's rule, or in this case Hooper's rule, the same thing in this case, says that the answer is one minus, one minus P, one times one minus P_2 and the reasoning behind that is easy to understand. This is the probability that the first one will be careless, this is the probability that the second one will be careless. If they agree the only way they can be wrong is if both were careless. This is the probability of that happening and this is the probability of that not happening.

So in general, this is the probability that at least one of them would be truthful. Well, now you have, that is working on one probability frame and then this sort of general rule I talk about transferring probabilities from one frame to another frame when you have a compatibility relation, seeing what their testimony is, seeing the fact that they agree on saying that it is icy outside. Seeing that creates a compatibility relation, that says if at least one of them was being truthful, then it is true that it is icy outside. So this would be a valid belief function argument.

For example, if you gave each witness separately only a credibility of three-fourths, together their testimony would carry weight 15/16, .9375. Okay, so we could give a Bayesian analysis of that story. We could say, well, this is not right, because it does not have the prior probabilities and everything in it. One way of explaining why it is not right is it does not take into account the fact that these two guys agree.

Let's suppose we made a Bayesian model. When we make a Bayesian model, we have to put in some additional judgments. We have got to decide in some way what the prior probabilities are and also what the probability of them being accurate is if he is being careless. Let's suppose for argument that if he is careless, when they are careless they are always wrong. In other words, if they are careless they lie to us somehow.

Well, in that case, you can calculate the prior probability by Bayesian principles and it comes out not 15/16 but 9/10. So it is a different answer. But again, I make the same point. If you could make these additional judgments, if you can construct a convincing argument that says you have evidence for this kind of a judgment for the probability that this guy is going to be accurate if he is being careless, then this is a better argument than the belief function argument.

But if you can't, then the belief function argument may be the best you can do. So for Professor Lindley's statement that the probability is always adequate, I think the answer has to be, well, that means that you can always write down a beautiful Bayesian analysis, which would show what we would like to do in the sense that we would like to have those inputs to make that argument convincing. But sometimes we don't have the evidence needed to support those judgments.

So there is my argument.

DR. DeGROOT: Thank you. Lotfi, do you want to make some comments please?

DR. ZADEH: I would like to focus my comments on one recurring issue that has been heard here and that is the issue of adequacy of probability theory and an issue that was very forcefully argued by Professor Lindley. It seems to me that there are really three points that can be raised in that connection.

Professor Lindley maintains in effect that numerical probability theory is adequate, the theory in which probabilities are treated as numbers. I think that Professors Dempster and Shafer took issue with that and said, no, we have to go beyond that, that we have to admit interval data probabilities and then they said, well, probably that is a good place to start, although they don't take as strong a position on that as Professor Lindley does.

My own position is this, that one in many cases has to go beyond that, in other words, beyond interval data probabilities. It is not that one should not use probability theory, but what kind of probabilities one should be allowed to use.

What I am saying, at least in part, is that one should be allowed to use linguistic probabilities, which are basically imprecise characterization probabilities. Now, the classical linguistic probability is a special case of the linguistic variable. What is a linguistic variable? Well, here is an example.

There is something that admits to a numerical characterization in this case, but that need not be the case, like some sort of numerical scale, tall. So we can describe it in numbers, but there are many situations in which we either do not know really what the number is or we will find it unnecessary to specify what that number is exactly. So it might be sufficient, as we frequently do in every day discourse, to say it is medium or to say it is low or to say it is very low. You have curves like that which are generalizations of intervals.

These generalizations of intervals are possibility distributions. Here is another slide which shows that perhaps more simply. Here we are talking about age. So you have young, you have old, you have not young, you have very young and so forth. Now in this sort of characterization, instead of using the constant of a unit, which is something like the canonical examples that we talked about here before, we are using two primary fuzzy sets, young and old.

Now, once these have been calibrated in a particular context, then you use these modifiers like <u>very</u>, <u>rather</u>, <u>quite</u>, <u>somewhat</u>, <u>extremely</u>, <u>more or less</u>, <u>not very</u> and so forth, to generate other values, and this is how it will work in the case of probabilities. So the primary terms are <u>likely</u> and its antonym, <u>unlikely</u>. Then you have <u>not likely</u>, <u>very likely</u>, <u>not very likely</u>, <u>more or less likely</u>, <u>extremely likely</u>; and on the other side you have the same sort of thing with

unlikely and here you have mixtures of these, not likely and not unlikely.

So what happens then is this, once you define or calibrate likely and unlikely, then from that point on the definitions of other terms can be computed automatically. In other words, each of these modifiers is interpreted as an operator. These operators then act on the primary sets and generate other sets. This is the basic idea behind linguistic variables. Notice that in this sort of a thing, you can replace likely by handsome and you would have handsome, ugly, not handsome, very handsome, not very handsome, more or less handsome, extremely handsome and so forth. Anything you can think of you can substitute in there and it would be the same sort of a thing.

The point I am trying to make is that in every day discourse we use this sort of a thing all of the time. We can and we do use the same sort of things with respect to probabilities. So what it boils down to really is this, that instead of trying to force people into the use of numerical probabilities, you allow them to use linguistic probabilities with the understanding, however, that these linguistic probabilities are labels for fuzzy sets.

So they are not treated like some labels that you can't really go inside. You can go inside of these things. You can make use of the more detailed structure of these things and you have a system for generating complex values out of simple values. In effect, you have a language. This is what we call the language with a semantic structure in the sense that given the, it has a syntax, and given the syntax tree for any one of these labels, we can find, we can compute then the meaning of the label.

So the point I am trying to make now here with the simple examples, if this is likely, unlikely is the mirror image of it, the antonym. Not unlikely is one minus that. More or less likely is interpreted as the square root of the number likely. Very likely is interpreted as a square and so forth. Whether it is square root or square is not important.

The essential point here is that it is some sort of an operator which modifies the possibility of distribution. So if you consider problems of the order of complexity of what David presented here this morning, then it seems to me it is necessary to both the patient and the doctor or diagnostician or clinician, whatever it is, to make use of characterizations of this kind in situations in which the use of numerical probabilities cannot be justified on the basis of the information that it will.

Now the same sort of a thing applies to the numerical context in diseases. What do you mean by arthritis? You cannot really come up with simple definitions of complex diseases. The same thing applies to, for example, recession. What do we mean by recession? You see at this point what people try to do. It is extremely simplistic, like if the gross national product decreases in two successive quarters, then we are in recession. But that does not capture the complexity of the concept.

So what we have to do then, is we have to consider various constituents of that concept, like GNP, unemployment, increase in bankruptcy and so forth and then have a table which tells us that the decline in GNP is little and unemployment is low and bankruptcy is high, then the degree to which we are in recession is not true and so forth.

What I am perhaps harping on is the idea that in dealing with complex issues, we are using at this point inadequate tools. We are simply not matching the complexity of these concepts with a system that allows us to capture some of this complexity. So this was my point.

DR. COHEN: Are you advocating the use of a table like that?

DR. ZADEH: Yes.

DR. COHEN: Doesn't it take the place of the use of the fuzzy set of recession? You can simply treat the values under GNP, unemployment and bankruptcy as evidence in the Bayesian update.

DR. ZADEH: This is a different issue.

DR. COHEN: I would think the use of the fuzzy set would be where you don't want, it is too complex to create a table like that and so you simply ask for suggestions, the degree of recession.

DR. ZADEH: Of course, you know in this presentation. I cannot go into the details. I am speaking of many things. But roughly what is involved is, for example, in the case of unemployment, low means this; more or less high means this and so forth, and then there is a formula. There is a formula which takes this kind of a table and it is called a translation formula, and translates that into a relation which is defined no longer on these labels.

And from that point on you can interpolate this table, so that if you have an entry like this is slightly over little and this is more or less low, something that is not in the table, then you can find the degree to which you are in recession. So basically it is a matter of it allows you to interpolate. You cannot interpolate if you stick with just labels, just as labels. That is what happens.

The same sort of thing happens in the case that David presented this morning. If you treat these labels as simple anatomical things, you cannot interpolate. You need many, many more rules. If you have the capability for interpolation, you can get by with fewer rules. Otherwise, you have to make a provision for every eventuality and that is impossible.

Let me then say just very briefly something about belief and plausibility and this is a point that I mentioned yesterday. I do have some objection to the use of the word belief and my objection is the following, that basically then because of the incompleteness of our information, we cannot put in a probability that has interval value. We have the lower bound and we have the upper bound, but attaching the name

belief to the lower bound, we tend to lose sight of the fact that this is simply the lower bound, that we are dealing in fact with an interval value probability and the user is misled into believing that this is all that the user needs, because the user is told that the belief is .3 and the user says that is enough. All I am interested in is the degree of belief. The user is not told that this is simply an interval of which that is the lower bound.

So the plausibility is somehow, although it is present in the theory, but somehow it is paid much less attention to and they tend to focus their attention only on one bound.

DR. SHAFER: I know I had my turn, but I do want to say that I don't regard the belief function thing as giving interval probabilities. Like Fred's testimony gives me 80 percent reason to believe it is icy outside and zero to believe it is not, I don't regard that as bound at 80 to 100. I just regard it as 80 on one side and zero on the other.

A bound implies that further evidence might give something more definite between 80 and 100. Further evidence might give something less than 80.

DR. DEMPSTER: That is my vote too.

DR. ZADEH: I realize that, but unfortunately I don't have the time to go into it. But I would be prepared to argue this point that we are dealing with interval value probabilities. That is all we are dealing with.

Well, I will stop at this point. Thank you.

DR. DeGROOT: Thank you.

DR. LINDLEY: One thing that has surprised me from this conference is the support that everybody has given to probability. From the readings of writers on belief functions and fuzzy logic, I had not gotten the impression that probability played a very important role, but it is clear, I think, for me at any rate, from the discussion that it does play an important role even in those approaches. And Professor Zadeh said yesterday some encouraging words about how he would use probability and Glenn Shafer used encouraging words about how he would use probability if he had enough information to do so.

So I feel there is a fairly solid base for using probability and what we are really discussing is whether it is going to work all of the time. If I understand the other people correctly, they are saying there are situations in which it would be nice to use the probability argument, but circumstances prevent it. That seems to me to be quite a bit of progress. It is for me.

Let me now look at these inadequacies. A remark of Art Dempster yesterday very much puzzled me. He said, I am afraid it surprised me so much I did not write down his exact words at the time, but he said something to the effect that belief functions were just a generalization of probability, that probability was known about all of the time and that it was an addition to them.

This does not seem to me to be right, because belief functions combine according to his own rules, not according to Bayes rules and the Great Scorer in the heavens above us will score him rather badly when he does this. The rules are different and that seems to me to be very important.

Something else too that Glenn Shafer said I take exception to, and I did take this one down. He said you can't take things out of a Bayesian argument and expect it to work. This was in connection with modularity. You can't take things out of a Bayesian argument and expect it to work. Now that seems to me to be very surprising, because that is one of the great strengths of the Bayesian argument, that you can indeed just do that.

Any of you who have done any statistics know that one of the great strengths of the Bayesian argument is that you can remove nuisance parameters without any difficulty at all. You just integrate them out and this is one of the great arguments. So if you don't want that thing or if you don't want that parameter, fine, you just integrate it out.

And so this argument, this statement did seem to me to be unsatisfactory.

Another feature too which does seem to me to be a little confusing is the role of frequency in the belief function argument. For example, today he talked about imbedding in a sequence of problems. The Bayesian argument has nothing whatsoever to do with imbedding in a sequence of problems. It is a one off judgment using the information that you have. The information may refer to some other problems, but it has nothing to do with it and this is the great dispute between the classical statisticians and the Bayesians.

Returning now to Professor Zadeh, he made one statement yesterday, that Bayesians assume that we can make up for incompleteness by subjective probability. I hope that you understand that that is incorrect. I at any rate do not assume this. Certainly to make that assumption seems to me to be gross. In fact, I might add a little bit of personal history,

I was taught statistics at Cambridge by what I think of as the leading Bayesian, Harold Jeffreys, and I just did not go along with Harold Jeffreys, because he asked me to believe that you would combine your judgments by the laws of probability and that frankly was too much for me to swallow and that was the reason I did not accept his argument.

Then along comes Savage and then I learned Ramsey had done this eefore and De Finetti and there is a convincing argument to it, so one does not assume these things. One in fact proves them. In fact, if you read Harold Jeffreys very carefully and a little bit charitably, I admit, you can see that he is in fact edging towards a proof that he really was not making this assumption in his stuff.

There was another statement made by Professor Zadeh, this one was on the screen, so I am sure I have got it right, what matters in decision analysis is usual rather than the expectation. I have been doing some work in connection with nuclear power stations and the usual thing with nuclear power stations is that they work and they produce electricity 99.9 percent of the time. But what really matters is what is going to happen on that very, very unusual circumstance when they don't work. In fact, all of the research goes into that activity. There is an example that that statement is certainly not correct.

He made some remarks, you may remember too, about car insurance, that you could not evaluate the probability that your car will be stolen. Now, this actually, this sort of thing happens quite a lot of the time. I have done quite a bit of consulting for an insurance company. If you are, and I think it is quite realistic, to be in the situation where you have difficulty in evaluating the probability that your car will be stolen, think about how much insurance you are prepared to pay. Because one can work out from the amount of insurance you are prepared to pay what the probability is. You don't have to ask people probability questions in order to find out what that probability is and observe, I don't know what the law is in the United States, it varies from state to state, if taking out car insurance is voluntary, people do decide whether to do it or not and they are tacitly making an assumption about that probability.

I was delighted, of course, with David Spiegelhalter's talk. But even he felt that probability was not quite adequate all of the time and, again, I am afraid I don't agree. For example he talked about the likelihood of the union of A_1 and A_2 given X and suggested that it was the maximum of the likelihood. This is not right. You cannot infer what the likelihood of A_1 union A_2 is entirely in terms of the separate likelihoods.

This was a point made by Fisher when he introduced likelihood. There is no formula involving maximum or anything else that will do the job for you.

He talked about Idiot Bayes and I agree with him it is a bit idiot. What, of course, we want to do is pukka* Bayes, but if we can't do pukka Bayes then we can do kuccha* Bayes.

---

*Pukka in Hindi means ripe; kuccha means raw.

One thing that nobody has addressed at this conference is the problem of decision making. I still do not know how the fuzzy folk or the believers make up their minds as to that.

I made a challenge that anything that could be done by these other methods could be done by probability. So I am just trying to beat the Bishop of Welles and Bath. I haven't much time to do it. Here is the Bishop of Bath and Welles and the Bayesians and we want to know whether event A is true or not, so we want to calculate the probability of event A. The evidence we have is a_1 and a_2. Witness one said it is so and witness two said so. But to repeat, a_1 and a_2 are the pieces of evidence from witness one and witness two and A is the event.

Witness one said it was true and witness two said it was true. So the Bayes argument begins by saying what is it you don't know? We don't know whether the event was true or not. What do we know? We know a_1 and a_2. So therefore on the left-hand side the thing I want to calculate is the odds on event A, given a_1 and a_2, and on the right-hand side, I have put it down in Bayes form.

Now, I have to do some calculations and the first thing that I realize is that there is nothing in the data that Glenn Shafer put on the screen to enable me to go any further. Because he did not tell me anything about the probability that these two witnesses would separately state a_1 and a_2. Perhaps he meant they were independent.

If they were independent, then I could do the following, provided I recognized that they were independent given the event is true and also given the event it was false. This reminds us that whenever we are considering evidence, we have to consider the evidence on the supposition of guilt and on the supposition of innocence. And it is extraordinary to me that a lot of the writing about witnesses carrying on from the Bishop of Bath and Welles failed to recognize that. They talk about the reliability of witnesses, as though it were one number. It is not.

The witness's reliability is two numbers - the probability that he would say this when it is true, and the probability that he will say the same thing when it is false. Both of those things are relevant and there are circumstances in which they can be entirely different. That is, the person could be extremely reliable when the event is true and extremely unreliable when the event is false.

So, therefore, the fact that the Bayes handle has produced this result seems to me to be one up for Bayes. Now, let's assume they are independent and if I do assume they are equally reliable on a not A, I get that result and that is the result that Shafer put on the board. So we now see that in order to get this result we have had to make two very important assumptions. The first assumption is independence and the second assumption is equal reliability.

DR. WISE: If you assume p_1 and p_2 are nine/tenths, doesn't it give a ratio of 80-81 and a probability of 81/82 instead of what you got?

DR. LINDLEY: I don't think it does.

DR. DeGROOT: Let's continue on. I think you can worry about that later.

DR. LINDLEY: But I had to make two assumptions here. What I say to you is this, don't you think that those two things are relevant? Don't you think it is relevant to think whether those two witnesses are independent or not? An argument that does not take account of that, do you feel happy with it? Do you really feel happy with not having to bother with whether those witnesses are independent? Do you really feel happy with not having to bother whether the event was true or false? Do you really feel not happy about not having to put P(A) in?

Suppose you knew A was almost certainly true, would you really want to discount it? Do you really feel happy? You see there are only two possibilities. Either the argument that Glenn produces agrees with the Bayesian argument or it does not. If it does not agree with the Bayesian argument then the great scorer will have it.

If it does agree with the Bayesian argument, then he is making some assumptions somewhere and I ask you are those assumptions reasonable? Now, I would not guarantee that this piece of calculation is correct. I am not very good at doing calculations quickly and I was trying to listen to Professor Zadeh at the same time, but it did appear to me when the calculations are done that Dempster and this independent Bayes will agree if this holds. That is a very curious statement. If the probability event is not true is equal to the probability that the witnesses will say a_1 and a_2.

This has nothing to do with these probabilities up here. This is the probability that they will say that a_1 and a_2 unconditionally. So if you are a Dempster, you are making that sort of statement. Is that really reasonable? Do you really believe that?

What I am saying to you is that if you do the probability argument in full, turning the mechanical handle of the calculation, it will show you there are certain things you have got to think about. Think about them. They are in this case, and I think you will always find, that they are relevant and an argument that does not use them, it seems to me, is very suspicious.

Thank you.

DR. DeGROOT: Thank you. David, do you want to take a couple of minutes?

DR SPIEGELHALTER: I want to talk purely on a practical level rather than arguing about the theory of any of these approaches. First, on the fuzziness, there are two levels in which Professor Zadeh has been saying fuzziness can be used. The top level of fuzziness has to do with whether the propositions themselves are ill-defined or not. I have a couple of pictures that I will use, that I stole from other people, to illustrate the difference between a well-defined proposition with sort of probability attached to it and an ill-defined proposition which has a degree of truth attached to it.

This one, is it fish or fowl? It sort of has fish on some faces and fowl on the other. They have a degree of fish and fowl.

Also another example I used for a proposition is that I can read the next overhead. Well, this is only partly true. I know that is a fuzzy something or other. I have no idea what it is.

DR. DeGROOT: Does that mean that fuzzy is untranslatable?

DR. SPIEGELHALTER: My argument was that it is both unsatisfactory to have fuzzy propositions and it can be avoided in general by identifying the propositions used by crispifying them in terms of the actual interaction that is on house with the system.

The second order of fuzziness is when we start saying we are going to use probabilities but we are not quite sure what they are. And should we in fact say it is low, around about .2 or something like that? Again, I would say from the examples used this morning that when we do ask people probabilities and they don't know what they are, if you sit them down and talk to them hard enough, they will give you some idea of ranges and draw curves and they will do it.

And then there are data probabilities. You might not even feel too happy about the curves they have drawn, but at least it tells you that, how on the input you need more information, how you can update those probabilities and you can learn about those probabilities.

So both in the top level and on the second level, I feel that fuzziness can be avoided and I would like to avoid it.

Then as to the argument on the practical thing about the belief functions, I am not going to argue about the theory and justification for it. What I went over this morning, what I wanted to argue, was that there are certain behavioral demands called belief functions from people designing expert systems. Specifically, it is because they want to work with hierarchies of hypothesis structures and a hierarchy of taxonomy. They want to deal with ignorance and one will say, well, we just don't know anything about the lower levels of this hierarchy and because belief functions provide the method of identifying sources of evidence explicitly and so you can identify what contribution each source of evidence is towards the final conclusion.

And all of these seem to be very reasonable demands, but my claim is that they all can be dealt with within a probability calculus. One has to put more in because one ideally wants to define a joint distribution over all possible propositions. And you are necessarily going to have to use all sorts of approximations and ways of padding out distributions.

But essentially what I went over this morning was designed to say that you can cope with hierarchies and you can cope with ignorance and in fact within a closed expert system and provide an operational definition of ignorance in terms of possible beliefs that you may have when further information comes in. One can identify sources of evidence through this rather crude but effective way of showing how individual events and evidence has changed your beliefs in the past.

Coming on to Professor Lindley's points, I have in fact changed my mind, I think, since this morning, since talking to him and Ben Wise. My feeling is that the limitations of probability are when you, for some reason, want to use ill-defined propositions or you want to use propositions that are not strictly verifiable.

I had previously thought that maybe, in cases, for reasons of control, you might want to use something that was not strictly probabilistic. If you talk to people from other areas and they talk about, well, you have got this situation where you want to decide whether to trigger a particular set of rules, a particular set of possible hypotheses, very much in the INTERNIST idea; or trying to develop a differential diagnosis, trying to structure your problem. In doing that structuring perhaps you might want to use ideas of relevance that a particular symptom makes you want to look at a particular disease. That idea of relevance could be related to whether that symptom, in some way, provides a description of that disease that gives some support to that disease. It might be in terms of using a calculus that supports what a set of data gives to a set of possible hypotheses. In fact maximum support to any particular member of the hypothesis and the maximization of a likelihood is looking like the sort of thing that is done; something comparing two groups and hypotheses in likelihood ratio tests, in which one does not maximize over the likelihood.

I have changed my mind since this morning. I don't think that is necessary and I believe one should be able to work within the probability calculus by, at any time you are considering extending your frame of concern and considering new hypotheses, that these should be brought in and the probability of distribution should be assessed on those hypotheses and a decision to pursue a particular line of reasoning can be based on a probability.

So in that way, I have become a bit more extreme during this discussion. I believe if you are working with theoretically, verifiable propositions, then one need only consider the probability.

DR. DeGROOT: On that happy note, I now reveal myself as a true believer I have sort of felt over the last couple of days like the Barbra Walters of the theory of uncertainty or something and we have now come down to the wire and I think that we should not leave until we have settled this issue. So are you ready to vote?

(laughter)

DR. DeGROOT: I want to thank the speakers and all of you for participating. We adjourn the session.

RETROSPECTIVE COMMENTS

Stephen R. Watson

RETROSPECTIVE COMMENTS ON PAPERS AND PRESENTATIONS

STEPHEN R. WATSON
Cambridge University

1. Introduction

These notes contain my comments as a discussant at the conference on the Calculus of Uncertainty in Artificial Intelligence and Expert Systems, which was held at George Washington University on 27 and 28 December 1984. In the next four sections I give an account of the points that I made at the end of the four main talks at the conference, by Professor Glenn Shafer, Professor Lotfi Zadeh, Professor Dennis Lindley and Dr. David Spiegelhalter. In section six I present some summary comments which were not made at the conference, but are made now as a result of my reflection on what was said at the conference.

2. Comments on the contribution of Professor Shafer

One of the things that makes Shafer's theory interesting, is that it can be seen as an alternative to the traditional probability theory. Is this really so, however? Firstly, note that one of the strengths of subjective probability theory, is the clear cut-nature of the axiomatic support for the theory. Indeed, as Professor Lindley's contribution showed, it is possible to claim that probability theory is the only theory one could possibly use to represent uncertainty. Shafer's theory does not as yet have such a clear-cut support. For example, although Shafer recognizes the importance of canonical examples, as yet belief function theory is not provided with the same axiomatic development that is available for probability theory.

It can be claimed, however, (see Dempster's contribution at this conference) that belief functions are indeed rooted in probability theory. It is just that the probability is associated with a power set rather than a simple set. If this interpretation of belief function theory is accepted, then indeed there is no problem because the philosophical support for probability theory clearly also will support belief function theory. However, Professor Shafer seems in some of his writings not to be very happy with this interpretation of his theory. And if he rejects this interpretation then the problem of a philosophical foundation for belief function theory remains.

The second point I make here concerns concepts of independence. Professor Shafer touched on this point in his talk, but it is worth saying again that concepts of independence in belief function theory are not yet clear. Firstly, in the application of Dempster's rule to determine the support for a hypothesis on the basis of two pieces of evidence, there is a rather vague notion that the two pieces of evidence should be independent in some way. The detailed meaning of this concept of independence is far from clear. Shafer recognizes this difficulty and in his discussion of frames is attempting to overcome it. It is sufficient to say at this point, however, that we do not yet know how to handle dependence concepts in belief function theory in a way which is

intuitively understandable.

3. Comments on the contribution of Professor Zadeh

Firstly, note that Zadeh sees fuzzy set theory as a companion to probability theory, not as a replacement for that theory. Thus he sees some situations in which the use of the probability calculus is appropriate, but others where it is inappropriate. Indeed he sees fuzzy set theory as a calculus for handling imprecise entities rather than uncertain entities. Imprecision is a property which scientists have for many years been keen to avoid; yet one of the main reasons for Zadeh's introduction of the concept of the fuzzy set, was his belief that in Systems Analysis a stress on precision was misleading. It is always possible for the probabilist to claim (as indeed Professor Lindley did during this conference) that in any context imprecision can be modeled using probability theory. For example, if you are imprecise in giving me some information I am uncertain about what is actually the case. If you tell me that John is tall, I am uncertain about his precise height. It is thus clearly possible to avoid the need to introduce fuzzy sets; the cost of doing so, however, is to produce an enormously complex probabilistic framework which may well be impossible to analyze. (I will return to this point when I discuss Professor Lindley's contribution). To seek, therefore, to handle imprecise concepts directly, rather than to introduce precision and accompanying uncertainty seems to me to be a virtuous aim. To the extent, therefore, that computations using the fuzzy set theory give sensible results, it seemed a useful heuristic to follow.

It must be admitted, however, that problems exist in using fuzzy set theory. Perhaps the most obvious of these is the origin of the numbers that go to make up a membership function. As I mentioned in section two Professor Shafer agrees with the probabilists, that one needs canonical examples in order to provide meaning for the mathematical constructs one uses. Such examples do not exist within fuzzy set theory. When taxed with this question (as indeed he was at this conference) Zadeh points out that people seem to have an intuitive idea of how to provide such numbers, and that if we bother too much about precisely what the numbers are, we vitiate the whole spirit of the approach which is concerned with the representation of imprecise quantities. But this does not answer the problem fully. The open question in my mind is how sensitive the outputs of fuzzy analyses are to the representation of imprecise concepts by membership functions. If outputs are indeed sensitive then the precise choice of a membership function is rather important, and at present there is no guidance within the literature on how to choose one membership function rather than another. On the other hand, if the answers are insensitive to these representations, then one wonders if the outputs of a fuzzy analysis can actually tell one anything.

Then again, there are the connectives. When Zadeh introduced fuzzy set theory in the first place, he suggested the max and min operators for the connectives 'or' and 'and' respectively. There are, however, a great number of other operators which could be used to represent these connectives, and have many of the same properties (such

as reducing to the traditional operators in the case of crisp sets). It appears at present that within fuzzy set theory, there is no protocol for determining which of these connectives to use; rather one is advised to use whichever seems sensible in any given context. This emphasizes fuzzy set theory as a heuristic approach. It must be thought of a a reasonable way to get sensible results in a complex analysis,. rather than a 'correct' approach following on from believable and acceptable axioms.

Finally, we should comment on the blandness of fuzzy set theory. Because it deals with possibilities rather than probabilities, it is quite easy to create input membership functions which are so broad (in the sense of allowing a large number of possible values to have nonzero possibility that the output fuzzy distributions are extremely broad. This is not surprising. The more imprecise inputs we put into analysis, the more imprecise we can expect the output of the analysis to be. I am not sure if this can be articulated into a general principle, since one can clearly construct examples where imprecision does not build on itself in this way. None the less, it is my impression that fuzziness can get out of hand. In such circumstances, of course, the solution is to go back to the beginning and be more precise, and a probabilistic analysis would demonstrate the need for this.

In summary then, I see fuzzy set theory as a sensible heuristic way of describing imprecise concepts, and of breaking through the complexities of other kinds of analysis. The fact that it is a heuristic, however, means that we can never be certain that the results of the analysis make sense.

4. Comments on Professor Lindley's contribution

The conviction with which Professor Lindley speaks, and the sheer power of his argument impel users of alternatives to probability theory to respond to his arguments. If we do not accept the inevitability of probability, why not?

Users of Shafer's theory or Zadeh's theory can, and in fact have in the past, respond that they do indeed accept the inevitability of probability. As Dempster has commented, belief function theory is founded on probability, and so there is no contradiction in using belief function theory at the same time as using probability theory. Moreover, as I have argued, one can think of fuzzy set theory as being a heuristic approach in situations where a full probabilistic analysis is far too complicated to be undertaken.

It is, however, also possible to take issue with Lindley's argument. In other words, it is possible to question some of the premises in his argument and thereby avoid the full power of his conclusions. Firstly, if one investigates the development of subjective probability theory exemplified by Savage's approach, it is possible to ask whether we are prepared to accept the axioms. It is a commonplace now that people do not behave as though they accept Savage's axioms, reasonable as they undoubtedly are. Of course, these axioms are normative and it can be argued, as indeed Lindley does argue, that the

fact that we fail to abide by the axioms does not mean that we should not attempt to do so. Indeed he would say that the first act of a rational man is to agree to the axioms, and then attempt to construct his behavior in accordance with these axioms. If, however, we are not prepared to do this, then what happens to us is a matter of practice. It could be argued that if we are consistent in our failure to abide by the axioms, then our opponents can turn us into a money pump or construct a Dutch Book of gambles against us. Of course, we do not do this in practice. We just recognize when we are about to get cornered in this way, and change one of our judgments, possibly in a yet more inconsistent way with our past judgments. There is, therefore, nothing mandatory about accepting Savage's axioms, and we can therefore escape Lindley's conclusions if we wish to.

In his contribution Professor Lindley gave a very clear account of an alternative way of showing the inevitability of the probability. This was based on the notion of scoring systems. It is indeed quite remarkable that no matter what kind of scoring system one adopts, the numbers that one employs to describe uncertainty must (after an appropriate transformation) satisfy the rules of probability theory. Compelling as this argument is, we have to point out that in practice no Great Scorer exists. There is nobody hovering about us being prepared to dock our pay, should we use numbers which fail to conform to the rules of probability theory in our descriptions of uncertainty. Thus while the argument is elegant and powerful, there is nothing inherently irrational in not accepting it, because in practice scoring systems do not exist.

Of course the proof of the pudding is in the eating. If it can be shown that in the long run any person who fails in his assessment of uncertainty to combine his numbers as though they were probabilities will lose out inexorably, then indeed we have a problem in refusing to accept probability theory. But to my understanding practical proofs of this kind are not yet available.

Thus it is possible to escape the inevitability of probability; it has to be admitted, however, that there is no alternative theory which has the strength of support, and elegant support at that, which is available for probability theory.

The chief drawback with using probability theory is the complexity that sometimes results, and the need to assess an often surprisingly large number of conditional probabilities. In legal work, for example, great difficulty can arise; some interesting work by Schum, at Rice University, shows how problematic probabilistic inference can get. In some simple murder case, with five pieces of evidence, he needed to make 27 probability assessments. Professor Lindley suggested the principle of Occam's razor should be applied to our topic: simplify where possible. Sometimes probabilistic analysis is far from simple.

5. Comments on the contribution of Dr. Spiegelhalter

Dr. Spiegelhalter's talk was a most interesting account of the construction of an expert system for medical diagnosis. In his talk he gave us some important insights into the practical problems of constructing an expert system, which was both computable and also useful. This raises the general question of how one determines whether a particular expert system, as represented on some computer, is actually a good one or not. The issues involved are very similar to those involved in validating a model. Firstly, one needs the system to be faithful to some normative principle. This entire conference has been concerned with the appropriate normative principle to use in representing uncertainty in expert systems, and in my view one should start with probability theory, but be prepared to adopt other approaches as heuristics or as richer representations of the issues involved. It seems that Spiegelhalter's approach has been similar.

Secondly, one could validate an expert system by its comparison with expert performance. One can ask whether the diagnosis achieved by Spiegelhalter's system was better or worse than that achieved by competent diagnosticians. There is of course a debate over whether an expert system should be compared in this way. Is the goal to reproduce the abilities of an expert, or to improve on the abilities of available human judges? If it is the former, then indeed it is sensible to compare performance with experts, but in this case one wonders why one should not use the experts themselves. This could be answered by observing that very often experts are in short supply. If, on the other hand, our goal is to improve on human inference behavior, then the criterion of conformity with some expert performance is not appropriate. A final measure of the appropriateness of an expert system is user satisfaction. To what extent do the people who interact with the expert system feel that the system is of use to them? In Spiegelhalter's case there are two kinds of people involved, namely the patients and the doctors. As Spiegelhalter observed, it is very important that the doctors are supportive of the endeavor, and do not feel that their professional competence is in any way being threatened. It is perhaps more important, however, that the patients feel that they are being properly attended to. Spiegelhalter seems to have achieved success on both fronts.

6. Summary comments

Although the purpose of the conference was to discuss the use of the different theories for the representation of uncertainty in expert systems, the principal speakers, perhaps wisely, devoted their discussion mainly to arguing the cases for the use of their different theories in general. On the basis of the discussions we had at this conference, it seems to me that one can summarize as follows. Probability theory has a strong intellectual support and in principle there is no reason why one should not be satisfied with this theory. It does, however, provide enormous problems of complexity and of independence judgments and as a matter of practice it is necessary to seek for approximations. Fuzzy set theory can be viewed as a heuristic for handling those situations where imprecise inputs and imprecise

inferences are required without the need to resort to the greater complexity of probability theory. Belief function theory can be thought of as a way of representing inferences from evidence within the probabilistic framework.

There are yet other alternative approaches to handling uncertain inferences which are not mentioned at the conference, and notable among these is the non-monotonic logic of Doyle. Recently Cohen (Cohen et al 1985) has suggested a combination of Doyle's theory, with both Shafer's and Zadeh's, which he has referred to as the non-monotonic probabilist. This seems an exciting possibility of approach to the problem at the heart of this conference.

Reference

Cohen, M. S., Watson, S. R., and Barrett, E., 'Alternative Theories of Inference in Expert Systems for Image Analysis', Technical Report 85-1, Decision Science Consortium, Falls Church, VA, January 1985.

SECURITY CLASSIFICATION OF THIS PAGE

AD-A163 642

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): GWU/IRRA/ Serial TR-86/2	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: The George Washington Univ. Inst. for Reliability & Risk Analysis; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: Office of Naval Research
6c. ADDRESS (City, State and ZIP Code): 707 22nd St., NW Washington, DC 20052	7b. ADDRESS (City, State and ZIP Code): 800 N. Quincy St. Arlington, VA 22217-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Office of Naval Research; 8b. OFFICE SYMBOL (If applicable): ONR-1111	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-85-G-0162
8c. ADDRESS (City, State and ZIP Code): 800 N. Quincy St. Arlington, VA 22217-5000	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61153N 14	4118	NR 4118-128 (NR347-128)	4118-128-01

11. TITLE (Include Security Classification) The Calculus of Uncertainty in Artificial Intelligence and Expert Systems (Unclass)

12. PERSONAL AUTHOR(S)
N.D. Singpurwalla, Principal Investigator; S.M. Selig, Coordinating Editor; See #19

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Final	FROM 1 Dec 84 TO 30 Nov 85	1986/01/15	268

16. SUPPLEMENTARY NOTATION

Proceedings of Conference held December 28-29, 1984

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	Artificial Intelligence Expert Systems, Uncertainty

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This is a collection of papers, presentations and discussions at a conference on dealing with uncertainty in artificial intelligence and expert systems. Three different approaches were examined, namely(1) The use of belief functions,(2) fuzzy set logic, and(3) probabilit[y]. A case study example of a probabilistic approach for expert systems in medicine was presented.' Authors of individual papers are G. Shafer, L.A. Zadeh, D.V. Lindley, and D.J. Spiegelhalter.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Edward J. Wegman	(202) 696-4310	ONR-1111

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

SECURITY CLASSIFICATION OF THIS PAGE

END

FILMED

3-86

DTIC